Contributions méthodologiques à l'estimation de la survie nette : comparaison des estimateurs et tests des hypothèses du modèle du taux en excès

Coraline Danieli

UNIVERSITE CLAUDE BERNARD - LYON 1

Contributions méthodologiques à l'estimation de la survie nette: comparaison des estimateurs et tests des hypothèses du modèle du taux en excès

THESE

Présentée et publiquement soutenue

Le 16 décembre 2014

Par Coraline DANIELI

Née le 12 Juillet 1987 à Paris XIV

Pour obtenir le grade de DOCTEUR de L'UNIVERSITE CLAUDE BERNARD - LYON 1

SPECIALITE : Biostatistiques

Directeur de thèse : Nadine BOSSARD

Co-directeur de thèse : Pascale GROSCLAUDE

Ecole doctorale E2M2 : Evolution Ecosystèmes Microbiologie Modélisation

Laboratoire de Biométrie et Biologie Evolutive - CNRS UMR 5558

Membres du Jury de la Thèse :

Mr Daniel COMMENGES, Rapporteur (DR - INSERM)
Mr Jean-Yves DAUXOIS, Examinateur (PU)
Mr Jacques ESTEVE, Invité (PU)
Mr Roch GIORGI, Examinateur (PU-PH)
Mme Christine LASSET, Examinateur (PU)
Mr Janez STARE, Rapporteur (PU)
Mme Nadine BOSSARD, Directeur de thèse (PH)
Mme Pascale GROSCLAUDE, Co-Directeur de thèse (PH)

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. François-Noël GILLY**
Vice-président du Conseil d'Administration	M. le Professeur Hamda BEN HADID
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur Philippe LALLE
Vice-président du Conseil Scientifique	M. le Professeur Germain GILLET
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est - Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud - Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme. la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. FLEURY
Département Chimie Biochimie	Directeur : Mme Caroline FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur Georges TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. Jean-Claude PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. P. FOURNIER
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

Remerciements

Je remercie tout d'abord la Ligue Nationale Contre le Cancer dont le soutien financier m'a permis d'effectuer ma thèse dans de bonnes conditions.

Je remercie Pascale Roy et René Ecochard pour m'avoir accueillie au sein de leur laboratoire.

Merci à Nadine Bossard d'avoir passé son HDR peut-être juste pour cette thèse. Je la remercie pour son suivi cadré et sa qualité humaine.

Merci à Pascale Grosclaude d'avoir accepté de co-encadrer ce travail.

Je remercie les rapporteurs de cette thèse, Daniel Commenges et Janez Stare pour avoir accepté de lire, commenter et juger ce travail.

Je remercie Roch Giorgi, Jean-Yves Dauxois et Christine Lasset pour avoir accepté de juger ce travail.

Merci à Aurélien Belot qui m'a initié dans le monde de la recherche. Je le remercie pour sa disponibilité, son soutien, sa bonne humeur et sa patience.

Merci à Laurent Remontet pour sa grande qualité pédagogique, scientifique et son sens critique, sans qui, je pense, le travail n'aurait pas autant avancé.

Merci à Laurent Roche pour sa rigueur scientifique, sa disponibilité et son soutien.

Merci à Zoé Uhry pour ses conseils et sa joie de vivre.

Merci à Jacques Estève pour son expertise statistique toujours aussi pertinente, qui nous confirme à chaque réunion que nous avons encore un long chemin à parcourir…

Je remercie l'ensemble du service de biostatistique, secrétaires, ingénieurs, médecins que j'ai le plaisir à retrouver régulièrement. Merci à mes co-bureaux ou plus généralement aux « jeunes » (certains se demanderont s'ils en font encore partie) pour cette très bonne ambiance et ces relations amicales qui se sont développées.

Merci aux groupes de travail Mesure/Censur/Yros, qui m'ont permis de travailler sur ce sujet fort intéressant et de développer des collaborations avec d'autres équipes de recherche, notamment l'équipe de Biostatistique de l'université de médecine de Ljubljana, Slovénie, dont font partie Janez Stare et Maja Pohar-Perme que je remercie de m'avoir accueillie dans leur département et pour l'intérêt qu'ils ont porté à mon travail.

Je remercie mon comité de pilotage (Roch Giorgi, Bernard Rachet, Marc Colonna) dont j'ai bénéficié des judicieux conseils à de plusieurs reprises.

Je remercie mes proches de m'avoir soutenu dans cet objectif.

Table des matières

Table des figures

Listes des tableaux

Notations

n	Taille de la population
$T_{c,i}$	Temps de décès dû au cancer pour l'individu i
$T_{\bar{c},i}$	Temps de décès dû aux autres-causes pour le patient i
$T_{a,i}$	Temps de décès attendu pour l'individu i
$T_i = \min(T_{c,i}, T_{a,i})$	Temps de suivi observé en l'absence de censure administrative pour l'individu i
C_i	Temps de censure administrative pour le patient i
$U_i = \min(T_i, C_i)$	Temps de suivi observé en présence de censure administrative pour l'individu i
$\lambda_{o,i}$	Taux de mortalité observé pour le patient i
$\lambda_{c,i}$	Taux de mortalité en excès pour le patient i
$\lambda_{\bar{c},i}$	Taux de mortalité autres-causes pour le patient i
$\lambda_{a,i}$	Taux de mortalité attendu pour le patient i (approximation du taux de mortalité autres-causes pour le patient i)
$\Lambda_{o,i}$	Taux de mortalité cumulé observé pour le patient i
$\Lambda_{c,i}$	Taux de mortalité cumulé en excès pour le patient i
$\Lambda_{a,i}$	Taux de mortalité cumulé attendu pour le patient i

$S_{o,i} = \mathrm{P}(T_i > t \mid z_i)$	Survie observe (décès toutes causes)
$S_{c,i} = \mathrm{P}(T_{c,i} > t \mid z_i)$	Survie nette (pas de décès autres-causes)
$\mathrm{S_{a,i}} = \mathrm{P}(T_{a,i} > t \mid x_i)$	Survie attendue (décès dus aux autres-causes)
z_i	Vecteur des covariables pronostiques pour l'individu *i*
x_i	Vecteur des covariables démographiques pour l'individu *i*
$N_i(t) = I(T_i \leq t,\ T_i \leq C_i)$	Processus de comptage pour le patient *i*
$Y_i(t) = I(T_i \geq t,\ T_i \geq C_i)$	Indicatrice d'être encore à risque au temps t pour le patient *i*
$N(t) = \sum_i N_i(t)$	Processus de comptage au temps *t* (nombre de décès au temps *t*)
$Y(t) = \sum_i Y_i(t)$	Nombre de patient encore à risque au temps *t*

Introduction

L'étude des durées de vie désigne l'étude du délai de la survenue d'un évènement au cours du temps. Différents types d'événement peuvent être considérés : le décès, l'apparition d'une maladie ou encore la survenue d'une récidive. Dans le cas où le décès correspond à l'évènement, la survie, qui désigne la probabilité de vivre au-delà d'un certain temps t, se révèle être un indicateur pertinent. La survie des patients est utilisée dans le cadre de la surveillance épidémiologique des maladies au niveau d'une population, mais aussi en recherche clinique pour l'évaluation des stratégies thérapeutiques ainsi que pour l'identification de facteurs pronostiques et la quantification de leurs effets. Pour étudier la survie d'un groupe d'individus, une information nécessaire pour chaque patient constituant l'échantillon d'étude est la variable aléatoire « Durée de vie », que l'on nommera T, qui représente le délai entre la date d'origine (généralement la date de diagnostic) et la date d'apparition de l'évènement. Cependant, cette variable aléatoire ne correspond pas toujours au délai jusqu'à l'évènement étudié pour tous les patients du fait que certains patients sont encore en vie au moment de l'analyse. Ces patients sont dits censurés, c'est-à-dire que la seule information disponible est que $T > t$, t étant la date de fin d'observation: nous avons ainsi des données dites "incomplètes". Les informations nécessaires pour chaque patient sont alors le délai entre la date d'origine et la date de dernière nouvelle, et l'indicatrice d'évènement δ, qui représente le statut vital du patient à la date de dernière nouvelle. Ces observations censurées sont la spécificité principale de ces études. Une autre spécificité est que la distribution de cette variable aléatoire n'est, de manière générale, pas symétrique. Du fait de ces deux particularités, elle ne peut donc pas être analysée de la même manière qu'une variable quantitative ordinaire et nécessite une méthodologie adaptée pour l'analyse.

La survie peut se décliner sous différentes formes théoriques, la plus simple étant la probabilité de survie « globale » (aussi appelée survie brute ou observée). La probabilité de survie globale au temps t représente la proportion de patients vivant au temps t depuis le diagnostic de la maladie d'intérêt. Les premiers modèles de survie ont été développés afin de modéliser ce concept de survie globale, de manière non-paramétrique tout d'abord (méthode actuarielle [Böhmer, 1912] et méthode de Kaplan-Meier [Kaplan, 1958]). Par la suite, la prise en compte de différents facteurs pronostiques susceptibles d'agir sur la survie a nécessité le développement de modèles paramétriques ou semi-paramétriques (modèle de Cox [Cox,

1972]). Pendant longtemps, ces modèles ont été utilisés en considérant l'effet des covariables d'intérêt incluses dans le modèle comme constant au cours du temps (taux relatif constant au cours du temps) et linéaire en fonction des valeurs prises par ces covariables (le logarithme du taux relatif est une fonction linéaire des covariables). Les développements successifs de ces modèles ont permis de s'affranchir de ces hypothèses parfois trop contraignantes et non réalistes, d'autant plus à long terme.

Pour mesurer le réel impact d'une maladie sur la survie d'une population, il faut tenir compte du fait que cette population est également soumise aux mortalités « autres causes », en particulier si la population étudiée est une population composée principalement de personnes âgées. C'est ainsi qu'est apparue le concept de survie « nette », indicateur théorique, qui représente la survie que l'on observerait dans un monde hypothétique où la seule cause de mortalité serait la maladie d'intérêt. Bien que non observable, cet indicateur est le seul permettant la comparaison entre pays ou entre périodes de diagnostic concernant l'efficacité ou l'amélioration du système de soin du fait que la mortalité « autres causes » soit éliminée. Les premiers modèles ayant pour but d'estimer la survie nette ont été développés de manière non-paramétriques [Ederer, 1959][Ederer, 1961][Hakulinen, 1982] puis de manière semi-paramétrique [Sasieni, 1996] ou paramétrique [Estève, 1990] en adaptant les modèles développés en survie globale à ce nouveau concept.

Parmi les nombreux champs d'application de l'étude des durées de vie, la cancérologie est un domaine majeur. Le cancer touche plus de 350 000 personnes par an en France [Binder-Foucard, 2013]. Il fait partie des priorités de santé publique comme ont pu le montrer le développement de nombreuses structures dédiées à la recherche dans ce domaine ainsi que les différents « Plans cancer » mis en place par le gouvernement. C'est dans ce contexte que s'inscrit le travail de cette thèse qui portera particulièrement sur l'estimation de la survie nette en cancérologie.

La première partie de ce mémoire rappelle les concepts généraux de l'analyse de la survie. Les différents mécanismes de censure susceptibles d'être rencontrés dans des données de survie seront tout d'abord introduits. Les différents concepts de survie, c'est-à-dire, la survie globale et la survie nette, seront ensuite exposés. Puis seront présentés les différentes approches utilisées pour l'estimation de la survie globale et de la survie nette.

La deuxième partie de cette thèse, porte sur la comparaison des méthodes d'estimation de la survie nette. Une méthode d'estimation de la survie nette est la survie « spécifique »: cette méthode consiste à ne considérer, comme évènement, que les décès dus à la pathologie

étudiée. Les autres décès ne sont pas considérés comme des évènements et sont censurés. Cependant, cette méthode nécessite de connaître la cause de décès. Elle reste donc d'usage limité car la cause exacte de décès est souvent difficile à recueillir, en particulier à long terme [Percy, 1981][Ashworth, 1991]. Pour remédier à ce problème, plusieurs méthodes ne nécessitant pas la cause de décès ont été développées. Elles reposent toutes sur le principe suivant : « la distribution du temps de décès lié aux autres causes que le cancer est supposée connue; elle s'obtient généralement à l'aide de données de mortalité de la population générale ». Une première approche, appelée survie « relative », repose sur le rapport entre la survie observée et la survie que l'on aurait pu attendre, en l'absence du cancer, dans la cohorte de patient au vu de leurs caractéristiques démographiques (généralement âge, sexe, année de diagnostic, département de résidence). Une seconde approche repose sur une modélisation du taux en excès qui suppose que le taux de mortalité observé dans les données résulte de la somme du taux de mortalité attendu (en l'absence de maladie) et du taux de mortalité en excès (associé à la maladie). Enfin, une troisième approche repose sur la pondération par l'inverse de la probabilité de censure. Publiée en 2012 [Pohar-Perme, 2012], cette méthode est venue révolutionner la méthodologie de l'estimation de la survie nette. L'article présentant cette méthode a également permis de clarifier différents points d'ombre, qui avaient déjà été soulevés auparavant [Estève, 1990], concernant l'estimation de la survie nette. Cette publication est à l'origine de la deuxième partie de cette thèse. En effet, après la publication de cette nouvelle méthode dans une revue théorique, il était indispensable de comparer les propriétés des différents estimateurs disponibles. Après avoir comparé, en partie, les méthodes de façon théorique, le **premier objectif** original de cette thèse a été de comparer les méthodes à l'aide d'une étude de simulation dont le plan d'expérience a permis d'étudier l'influence de certains facteurs tels que l'existence d'un effet de l'âge sur le taux de mortalité dû au cancer ou l'existence d'un effet de l'âge sur le temps potentiel de suivi. Les résultats de cette étude ont donné lieu à une publication [Danieli, 2012]. Le **second objectif** original a été ensuite d'illustrer les résultats issus de ces différentes méthodes sur des données réelles issues de la base commune du réseau des registres français du cancer (base FRANCIM). Ces résultats ont donné lieu à une publication [Roche, 2013].

La troisième partie de cette thèse porte sur l'étude des outils diagnostiques permettant de vérifier que les hypothèses faites lors de l'utilisation d'un modèle paramétrique du taux de mortalité en excès sont correctes. Différents outils diagnostiques ont été proposés dans le cadre de la survie globale pour vérifier les différentes hypothèses d'un modèle, les plus

importantes étant l'hypothèse des taux proportionnels et l'hypothèse de linéarité ; en revanche, peu d'outils ont été proposés dans le cadre de la survie nette. La plupart de ces outils reposent sur les résidus, quantité représentant la différence entre ce qui est observé et ce que l'on attend du modèle. La difficulté dans l'interprétation des résultats (souvent graphiques) que fournissent certains résidus est que l'on ne sait pas toujours évaluer si le modèle est adéquat aux données ou pas. Une tendance observée peut être due à une mauvaise spécification du modèle mais peut n'être également que le fruit d'une variation aléatoire. Pour pallier à ce problème, une possibilité est alors de travailler avec des résidus dont la distribution sous l'hypothèse nulle H_0 est connue (H_0 : le modèle est correct) ; ce type de résidus s'appuie généralement sur les processus stochastiques. Dans cette troisième partie seront présentés tout d'abord les outils diagnostiques existants dans le cadre de la survie globale puis de la survie nette. A la suite de cet état des lieux, le **troisième objectif** original de cette thèse a été de développer des tests permettant de vérifier les différentes hypothèses d'un modèle paramétrique du taux en excès, dans le même cadre théorique des processus stochastiques. Le développement théorique a été fait pour évaluer l'hypothèse des taux proportionnels, la forme fonctionnelle ainsi que la fonction de lien. Cependant, les résultats de performance présentés ne concerneront que l'évaluation des taux proportionnels, les résultats des autres méthodes n'étant pas encore disponibles. A noter que pour ce chapitre, le lecteur doit être initié à la théorie des processus de comptage.

Ce travail de thèse a été conduit au sein de l'équipe Biostatistiques-Santé , équipe du Laboratoire de Biométrie et Biologie Evolutive - UMR CNRS 5558. Cette équipe a la particularité d'être adossée au Service de Biostatistique des Hospices Civils de Lyon, structure de soutien en biostatistique, opérant pour les équipes hospitalières mais également pour différentes institutions publiques ou privées. Il s'agit donc d'un environnement associant une facette « recherche en méthodologie statistique » mais aussi une facette riche en applications pratiques de l'outil biostatistique, et ceci dans tous les domaines cliniques. Le service est notamment impliqué dans un partenariat le liant aux agences sanitaires (Institut National du Cancer, Institut de Veille Sanitaire), et au réseau des registres français des cancers (réseau Francim). Dans le cadre de ce partenariat, le service est en charge de travaux d'analyse dont le cœur est l'épidémiologie descriptive des cancers. Il produit donc régulièrement des indicateurs épidémiologiques qui sont ensuite interprétés par les épidémiologistes du réseau Francim. Ceci conduit à des échanges permanents avec ces épidémiologistes, et de ces échanges surgissent régulièrement des questions nécessitant des

réflexions supplémentaires sur ces indicateurs ou des développements de nouveaux indicateurs. Une question posée de plus en plus fréquemment par les épidémiologistes et les cliniciens portent sur l'intérêt de la survie dite « conditionnelle ». Associée à la survie brute ou à la survie nette, la survie conditionnelle apparaît comme un indicateur complémentaire et pertinent, mais cependant peu répandu. Elle correspond à la probabilité de survie à y années après le diagnostic sachant que l'on a déjà survécu à x années ($y > x$). L'estimation de la probabilité conditionnelle de survie (brute ou nette) revient à estimer la courbe d'évolution (ou dynamique) du taux de mortalité ainsi que l'impact des facteurs pronostiques sur cette courbe. Le **quatrième objectif** de cette thèse est d'étudier l'impact de certains facteurs pronostiques sur la survie afin d'évaluer si, connus au moment du diagnostic, ces facteurs ont encore un effet, par exemple, 5 ans après. Autrement dit, « Comment l'effet du facteur considéré évolue-t-il au cours du temps ?». Cela rejoint sur le plan méthodologique la nécessité de bien estimer l'impact des facteurs pronostiques sur la dynamique du taux de mortalité en excès. Dans ce travail, nous nous intéresserons particulièrement à l'effet du stade au diagnostic : « Avoir été diagnostiqué en stade précoce ou en stade avancé a-t-il le même impact sur la survie selon que l'on se place au moment du diagnostic ou après avoir déjà survécu un certain nombre d'année, par exemple, 5 ans ?». Cette étude sera appliquée au cancer du côlon dont les données sont issues de données Hautes Résolutions (HR).

Enfin, la cinquième partie finit par les conclusions et perspectives de ce travail de thèse.

Chapitre I

Analyse des données de survie : Rappel

I.1. Mécanismes de censure

Comme énoncé dans l'introduction, l'une des principales particularités des études de survie est la présence de données censurées. Le type de censure le plus couramment rencontré est la *censure à droite*. Les patients censurés à droite peuvent être soit des exclus vivants qui sont encore en vie à la date de point (date fixée signifiant la fin du suivi pour l'analyse), qui n'ont pas vécu l'évènement, soit des perdus de vue qui sont sortis de l'étude et dont on ne connait pas le statut vital à la date de point. D'autres types de censures existent, tels que la *censure à gauche* lorsque l'évènement intervient avant la date d'origine de l'étude ou la *censure par intervalle* lorsque le recueil de données est fait à intervalle de temps fixé ; l'évènement survient dans un intervalle particulier mais nous n'avons pas connaissance de la date exacte. Ces deux derniers types de censures ne seront pas abordés dans la suite, nous nous focaliserons sur la censure à droite.

Les principaux mécanismes de censure à droite sont :

- la censure administrative ou date de point : Il s'agit de la date pour laquelle on cherchera à connaître l'état de chaque patient et au-delà de laquelle on ne tiendra pas compte des informations: en considérant l'individu i, nous pouvons observer le décès du patient i au temps T_i seulement si T_i est antérieure ou identique au temps C, temps écoulé depuis le diagnostic jusqu'à une date de point définie pour l'analyse. Autrement, la seule information disponible est que T_i est supérieur à C. On associe au patient i le couple (T_i, δ_i) tel que (T_i, $\delta_i = 1$) si le patient est décédé au temps T_i et ($T_i = C$, $\delta_i = 0$) si le patient est exclu-vivant et censuré en C.
- les perdus de vue : L'individu i est perdu de vue s'il sort de l'étude avant la date de point. La seule information disponible pour ce patient est alors sa date de dernière nouvelle. Il a été observé entre la date de diagnostic et le temps T_i. Le couple qui lui sera attribué est (T_i, $\delta_i = 0$).

- la présence des risques concurrents : Un individu peut être considéré comme censuré lorsqu'un autre évènement empêche la survenue de l'évènement d'intérêt. Par exemple, en se plaçant dans le cadre où notre évènement d'intérêt pour la patient *i* est le décès par cancer au temps $T_{c,i}$, le décès dû aux autres-causes peut empêcher la survenue du décès dû au cancer si $T_{a,i} < T_{c,i}$, $T_{a,i}$ correspondant au temps de décès dû aux autres causes pour le patient *i*. Dans ce cas, le temps de mortalité dû aux autres-causes représente le temps de censure du patient *i* par rapport à l'évènement d'intérêt qu'est le décès par cancer, ($T_{a,i}$, $\delta_i = 0$).

L'hypothèse d'indépendance entre les deux variables aléatoires, le temps d'évènement du patient et le temps de censure du patient, est très importante et permet une estimation non biaisée de la survie car elle suppose que les patients censurés n'ont pas une caractéristique particulière par rapport aux patients qui ne le sont pas. Cette hypothèse parait assez naturelle pour la censure administrative, dont on a vu qu'elle était déterminée par une date de point définie pour l'analyse et indépendante des caractéristiques des patients. Elle l'est beaucoup moins pour les deux autres types de censure, particulièrement pour les risques concurrents. En effet, en reprenant l'exemple, si le temps de survie dû au cancer et le temps de survie dû aux autres causes sont dépendants, c'est-à-dire, s'ils sont affectés par une même covariable, supposons l'âge, le fait de censurer les décès dus aux autres causes induit une censure que l'on nommera « informative ». En effet, ceux qui décèdent des autres causes et qui sont censurées sont particulièrement des personnes âgées. Celles-ci sont également les plus à risque de décéder du cancer. La censure devient non aléatoire. Elle doit donc être prise en compte lors de l'estimation de la survie. C'est une des difficultés qui sera rencontrée par les méthodes d'estimation de la survie nette.

I.2. Différents concepts de survie

I.2.1 Survie globale

La survie globale (brute ou observée) au temps *t* représente la proportion de survivants au temps *t* après la date de diagnostic quelle que soit la cause de décès (cancer ou autre). Elle est donc influencée par la mortalité « autres causes », en particulier si la population est composée de personnes âgées. Cette méthode ne reflète pas la mortalité due uniquement au

cancer et ne permet donc pas de mesurer l'impact de la maladie sur la survie de la population. Par exemple, supposons que la survie globale observée à 5 ans chez les patients diagnostiqués en 1990 est de 25% et de 50 % chez les patients diagnostiqués en 2000. Ces résultats ne nous permettent pas de savoir d'où vient l'augmentation de la survie ; aucune indication ne nous est donnée concernant la part que représentent les décès par cancer et les décès autres causes ni pour l'année 1990, ni pour l'année 2000. C'est ainsi qu'est apparu le concept de survie nette.

I.2.2 Survie nette

La survie nette représente la survie que l'on observerait dans la situation hypothétique où la seule cause de mortalité possible serait le cancer étudié. Cet indicateur est capable de restituer l'impact propre de la maladie. En éliminant les autres forces de mortalité, il permet la comparaison entre pays ou entre périodes de diagnostic. En effet, la survie nette n'est pas influencée par les différences pouvant exister entre pays ou entre périodes de diagnostic concernant la mortalité autres causes et permet donc une comparaison directe de l'efficacité et de l'amélioration des systèmes de soins concernant le cancer. En reprenant l'exemple précédent, les chiffres disponibles concernent maintenant la survie globale et la survie nette. A 5 ans de suivi, la survie globale et la survie nette était respectivement de 25% et de 75% pour les patients diagnostiqués en 1990 et de 50% et 80% pour les patients diagnostiqués en 2000. Contrairement aux résultats de la survie globale, les résultats de la survie nette permettent d'affirmer que la mortalité due au cancer a diminué entre 1990 et 2000.

Les différentes méthodes introduites pour estimer la survie nette sont présentées dans le chapitre II.

I.3. Estimation de la survie globale

Selon l'objectif de l'analyse de survie, différents types d'approches ont été introduites. Les sous-sections suivantes se focaliseront sur la survie globale mais ces différents types d'estimation de la survie existent également en survie nette, indicateur sur lequel nous nous focaliserons dans la suite.

I.3.1. Approche par modèle de survie non-paramétrique

Lorsque l'objectif de l'analyse de survie est uniquement une estimation ponctuelle de la survie (et qu'aucune information concernant les facteurs pronostiques n'est recherchée), l'utilisation de méthodes non-paramétriques pour estimer la survie suffit. Aucune hypothèse n'est faite sur la distribution des temps de décès. Les deux méthodes présentées dans la suite sont particulièrement connues.

Méthode actuarielle

La méthode actuarielle [Böhmer, 1912] diffère de la méthode de Kaplan-Meier par le fait que les intervalles de temps ne sont pas déterminés par la survenue d'événements, mais sont fixés à l'avance. On peut prendre par exemple des intervalles de temps égaux d'un mois, d'un trimestre, d'une année suivant le phénomène étudié. On estime alors la survie conditionnelle à la fin de chaque intervalle pour ensuite les cumuler pour l'estimation de la survie entre l'origine et le temps d'intérêt. Dans la suite, nous prendrons l'année comme intervalle de temps.

Soit n_i représentant le nombre de sujets en vie au début de l'intervalle de temps considéré i
Soit d_i représentant le nombre de décès observés pendant l'intervalle de temps considéré i
Soit w_i représentant le nombre de patients censurés pendant l'intervalle de temps considéré i

En absence de censure, la survie conditionnelle au temps t_i, s_i, chez les patients ayant déjà survécu au temps t_{i-1} est estimé par :

$$\hat{s}_i = \frac{n_i - d_i}{n_i}$$

En présence de censure, on suppose que l'occurrence de la censure est uniforme dans l'intervalle de suivi. L'effectif corrigé n' s'exprime alors de la manière suivante :

$$n_i' = n_i - \frac{c_i}{2}$$

Ce qui implique :

$$s_i = \frac{n_i' - d_i}{n_i'}$$

En cumulant les survies conditionnelles depuis la date de diagnostic jusqu'au temps t_i, l'estimation de la survie au temps t_i est la suivante :

$$\hat{S}(t_i) = \prod_{j=1}^{i} \hat{s}_i$$

Méthode de Kaplan-Meier

La méthode de Kaplan-Meier [Kaplan-Meier, 1958] « découpe » le temps de participation observé en plusieurs intervalles de temps ($[t_{i-1}$; $t_i[$ pour le $i^{ème}$ intervalle, t_{i-1} étant la date du $(i-1)^{ème}$ décès et t_i la date du $i^{ème}$ décès) afin d'estimer la survie sur ces différents intervalles.

Soit n_{i-1} le nombre de sujets en vie au début de l'intervalle considéré i, c'est-à-dire en t_{i-1}
Soit d_i le nombre de sujet décédés au temps t_i
Soit c_i le nombre de sujets censurés au temps t_i
Soit, par convention, $t_0 = 0$

La survie conditionnelle au temps t_i, s_i, chez les patients ayant déjà survécu au temps t_{i-1} est estimée par :

$$\hat{s}_i = 1 - \frac{d_i}{n_{i-1}}$$

En cumulant les survies conditionnelles depuis la date de diagnostic jusqu'au temps t_i, l'estimation de la survie au temps t_i est la suivante :

$$\hat{S}(t_i) = \prod_n \left(1 - \frac{d_i}{n_{i-1}}\right)$$

avec $S(t_0) = 1$ et $n_{i-1} = n_{i-2} - d_{i-1} - c_{i-1}$

I.3.2. Approche par modèle de survie semi-paramétrique

Les méthodes d'estimations semi-paramétriques ont été introduites dans le but d'estimer la survie tout en quantifiant l'effet pronostique de certaines covariables sans donner de forme particulière à la distribution des temps de décès. Ces méthodes sont représentées principalement par le modèle de Cox [Cox, 1972] qui s'exprime de la manière suivante :

$$\lambda(t, z) = \lambda_0(t)\exp(\beta z)$$

où λ(t) représente le taux de mortalité au temps t, λ_0(t) représente le taux de mortalité de base au temps t, $z = (z_1,\ldots,z_p)$ représente le vecteur des covariables agissant sur le taux de mortalité et β le vecteur coefficient obtenu à l'aide de la maximisation de la vraisemblance partielle [Cox, 1975] (partielle car le terme λ_0 est éliminé de l'expression de la vraisemblance car considéré comme un paramètre nuisible).

Cette vraisemblance partielle repose sur le fait qu'aucune information ne peut être donnée sur les coefficients aux temps pour lesquels il n'y a pas de décès. En ne raisonnant que sur les r patients décédés, la vraisemblance partielle individuelle du patient j qui correspond à la probabilité conditionnelle que cela soit le patient j qui décède au temps t_j, sachant les personnes à risque au temps t_j, est :

$$\upsilon_j(\beta) = P\left(D_j \middle| R_j\right) = \frac{\lambda_0(t_j)\exp(\beta z_j)\times dt}{\sum_{l\in R_j}\lambda_0(t_j)\exp(\beta z_l)\times dt} = \frac{\exp(\beta z_j)}{\sum_{l\in R_j}\exp(\beta z_l)}$$

où D_j représente l'évènement décès pour le patient j et R_j représente le nombre de patients à risque au temps t_j.

La vraisemblance partielle globale s'exprime alors de la manière suivante :

$$V^*(\beta) = \prod_{i=1}^{r} \upsilon_i(\beta) = \prod_{i=1}^{r} \frac{\exp(\beta z_i)}{\sum_{j \in R_i} \exp(\beta z_j)}$$

Après estimation des coefficients par le maximum de la vraisemblance partielle, la survie s'estime en utilisant la relation usuelle entre le taux de mortalité et la survie.

I.3.3. Approche par modèle de survie paramétrique

Les modèles de survie paramétrique supposent que les temps de décès suivent une distribution particulière ; ils permettent également de quantifier l'effet pronostique des covariables qui peuvent y être incluses. Des distributions classiquement utilisées sont : le modèle exponentiel, le modèle de Weibull, le modèle log-normal, le modèle gamma, le modèle log-logistique (Table I.1).

Table. I.1. Cinq modèles paramétriques usuels pour modéliser le taux de mortalité

Modèle	**Paramètres**	**Taux de mortalité**	**Forme**
Exponentiel	λ_0	$\lambda(t) = \lambda_0$	Constante
Weibull	λ_0, γ	$\lambda(t, \lambda_0, \gamma) = \lambda_0 \gamma t^{\gamma-1}$	Monotone
Log-Normal	μ, σ	$\lambda(t, \mu, \sigma) = \frac{\frac{1}{\sigma t}\varphi\left(\frac{\ln(t)-\mu}{\sigma}\right)}{1-\Phi\left(\frac{\ln(t)-\mu}{\sigma}\right)}$	$\cap$
Gamma	θ, ν	$\lambda(t, \theta, \nu) = \frac{\theta^{\nu}\Gamma(\nu)t^{\nu-1}e^{-\theta t}}{\frac{1}{\Gamma(\nu)}\int_0^{\theta t} u^{\nu-1}e^{-u}du}$	Monotone
Log-logistique	λ_0, γ	$\lambda(t, \lambda_0, \gamma) = \frac{(\lambda_0\gamma)(\lambda_0\gamma)^{\gamma-1}}{1+(\lambda_0 t)^{\gamma}}$	$\cap$

Les estimations des paramètres du modèle sont obtenues en maximisant la log-vraisemblance des observations (par l'intermédiaire de méthodes itératives, par exemple l'algorithme de Newton-Raphson).

Dans le cas des modèles de survie paramétrique, nous pouvons décomposer la vraisemblance V selon la contribution d'un patient vivant ($\delta = 0$) ou décédé ($\delta = 1$) :

- Pour un patient vivant au temps t_i, sa contribution à la vraisemblance correspond à la probabilité de survie au-delà de t_i : $S(t_i)$
- Pour un patient décédé au temps t_i, sa contribution à la vraisemblance correspond au produit de la probabilité de survie $S(t_i)$ par le taux de mortalité $\lambda(t_i)$

$$V = \prod_{i=1}^{N} S(t_i).\lambda(t_i)^{\delta_i}$$

$$V = \prod_{i=1}^{N} \exp(-\Lambda(t_i)).\lambda(t_i)^{\delta_i} \quad \Rightarrow \quad \log(V) = \sum_{i=1}^{N} -\Lambda(t_i, z) + \delta_i \ln[\lambda(t_i, z)]$$

Après avoir estimé les coefficients, la survie s'estime en utilisant la relation usuelle entre le taux de mortalité et la survie.

Chapitre II

Comparaison des différentes méthodes d'estimation de la survie nette

II.1. Description des différentes méthodes existantes

L'une des principales difficultés dans l'estimation de la survie nette est la présence de risques concurrents, particulièrement les décès « autres-causes » qui peuvent empêcher la survenue des décès par cancer. Ceci peut ainsi induire une censure informative si le temps de décès dû au cancer et le temps de décès dû aux autres causes sont dépendants. Nous verrons que cette censure informative n'est pas toujours prise en compte par les différentes méthodes d'estimation de la survie nette.

II.1.1. Cause de décès connue : méthode de la survie spécifique

Dans ce contexte, la cause de décès est supposée connue : l'évènement d'intérêt étant le décès dû au cancer, le décès dû aux autres causes est alors considéré comme une censure. L'inconvénient est que cette méthode nécessite de connaitre la cause de décès, ce qui se révèle être difficile du fait que les causes de décès peuvent être inconnues, inexactes ou incertaines [Percy, 1981][Ashworth, 1991] surtout à distances du diagnostic. De plus, le recueil de la cause de décès dépend fortement de l'équipe chargée de collecter et de restituer cette information. Ainsi, la survie nette estimée par cette méthode peut s'avérer difficilement comparable entre différentes zones géographiques d'un même pays, entre différents pays ou entre différentes périodes de diagnostic. Ce problème concernant la connaissance de la cause de décès existe également en recherche clinique et notamment dans les essais thérapeutiques où la cause de décès est généralement recherchée de façon standardisée. Cependant, même en considérant que la cause de décès soit fiable et parfaitement connue, un biais, assimilable à un biais de sélection, peut exister dans l'estimation de la survie nette par des méthodes de survie spécifique. Si les deux causes de mortalité sont affectées par les mêmes covariables (variables démographiques), c'est-à-dire, si les deux causes de mortalité sont dépendantes, la censure

due aux autres causes est « informative » et donc non aléatoire. Dans la réalité, le temps de décès dû au cancer dépend presque toujours d'au moins une variable démographique, l'âge; l'utilisation de cette méthode pour estimer la survie nette fournit donc des résultats biaisés si, à minima, l'âge n'est pas pris en compte. Classiquement, la méthode de Kaplan-Meier est utilisée en considérant les décès autres-causes comme des censures.

II.1.2. Cause de décès inconnue

Les différentes approches développées ici pour estimer la survie nette reposent toutes sur l'hypothèse que le taux de mortalité autres-causes d'un groupe de patients atteints d'un cancer puisse être approximé par le taux de mortalité de la population générale obtenu à partir des tables de mortalité selon l'âge, le sexe, l'année de diagnostic et le département de diagnostic du patient.

II.1.2.a. Méthodes dites « Ratio-Estimate »

Parmi ces méthodes se trouvent les méthodes de « survie relative », méthodes non-paramétriques, qui comparent la survie observée au temps t d'un groupe de patient atteints d'un cancer, $S_o(t)$, avec la survie attendue au temps t, $S_a(t)$, qui représente la survie d'un groupe de patients sains possédant les mêmes caractéristiques démographiques que le groupe de patients malades. La survie relative au temps t, $S_c(t)$, se définit comme le rapport entre la survie observée et la survie attendue [Berkson, 1950] :

$$S_c(t) = \frac{S_o(t)}{S_a(t)}$$

Plusieurs méthodes de survie relative, dont les différences portent sur la définition de la survie attendue, ont été développées afin de tenter d'approcher au mieux la survie nette. Par exemple, si la population initiale comprend 20% de patients appartenant à la classe d'âge la plus jeune et 20% de patients appartenant à la classe d'âge la plus âgées, après dix ans de suivi, les proportions ne seront plus les mêmes étant donné que les patients âgés seront

d'avantage décédés que les patients les plus jeunes : la proportion de patients appartenant à la classe d'âge la plus jeune sera donc supérieure à 20% alors que celle des patients appartenant à la classe d'âge la plus âgée sera inférieure à 20%. Les différentes méthodes tentent de prendre en compte la distorsion de cette proportion au cours du temps ; s'ils ne sont pas pris en compte, la survie attendue estimée à 10 ans de suivi sera alors sous-estimée.

Historiquement, les premiers auteurs introduisant la survie relative n'ont pas explicitement exprimé que les méthodes d'estimation de la survie relative avaient été développées pour estimer la survie nette. Avec le temps, le terme de survie relative s'est imposé comme un concept (au même titre que la survie globale ou la survie nette) alors qu'elle était une méthode proposée pour estimer la survie nette.

Méthode d'Ederer I

La méthode d'Ederer I [Ederer, 1961] est la méthode la plus « naturelle » pour calculer la survie attendue d'un groupe et est l'application directe de la table de mortalité. Elle repose simplement sur les probabilités conditionnelles. Il s'agit de la seule méthode étant réellement une méthode dite « ratio-estimate » car il s'agit bien du rapport de deux survies clairement définissables.

Pour un groupe de patient de taille n, la survie attendue du groupe depuis la date de diagnostic jusqu'au temps t, $S_a(t)$, est obtenue en faisant la moyenne des survies attendues individuelles jusqu'au temps t ; elle s'écrit de la façon suivante :

$$S_a(t) = \sum_{i=1}^{n} \frac{S_{ai}(t)}{n}$$

où $S_{ai}(t)$ est la survie attendue du patient i au temps t, survie déterminée à partir des tables de mortalité de la population générale, en fonction des caractéristiques (âge, sexe, année de diagnostic, département de résidence) du patient i.

La caractéristique principale de cette méthode est qu'elle estime la survie attendue à partir de la composition initiale du groupe et qu'elle ne tient pas compte du changement de structure de la population au cours du temps. Par exemple, un patient décédé à 1 an après le

diagnostic participe tout de même à l'estimation de la survie attendue jusqu'au temps *t*, même si *t* est supérieur à 1 an.

Bien que très anciennes, cette méthode a été utilisée pendant des décennies jusqu'à très récemment dans des études nationales et internationales (estimations nationales aux Etats-Unis - Seer - jusqu'en 2010 [Altekruse, 2010]).

Méthode de Hakulinen

La méthode de Hakulinen [Hakulinen, 1982] permet de prendre en compte le temps potentiel de suivi des patients. Le temps potentiel de suivi d'un patient représente le temps maximum observable pour le patient considéré, c'est-à-dire, la différence entre la date de point et la date de diagnostic. La méthode calcule la survie attendue des sujets susceptibles d'être observés au temps *t*.

$$S_a(t) = \frac{\sum_{i=1}^{n} C_i(t) S_{ai}(t)}{\sum_{i=1}^{n} C_i(t)}$$

avec $C_i(t)$ étant une indicatrice égale à 1 si *t* est inférieur au temps potentiel de suivi du patient *i* et 0 sinon.

Cette méthode permet donc de prendre en compte des phénomènes de censure tels que des situations où les personnes censurées n'ont pas la même survie attendue que les personnes non censurées. Cela peut être par exemple le cas si les personnes âgées sont inclues plus tard dans l'étude par rapport aux personnes jeunes. Les personnes âgées auront donc plus de chances d'être censurées que les personnes jeunes. Lorsque ce phénomène de censure n'existe pas, la méthode de Hakulinen équivaut à la méthode d'Ederer I.

Tout comme la méthode d'Ederer I, cette méthode a été utilisée jusqu'à très récemment dans des études nationales et internationales (estimations européennes Eurocare IV [De Angelis, 2009]).

Méthode d'Ederer II

A la différence de la méthode d'Ederer I, la méthode d'Ederer II [Ederer, 1959] met à jour le nombre de patients à risque à chaque intervalle de temps. En reprenant l'exemple précédant, le patient décédant à un an ne participe plus au calcul de la survie attendue au-delà de un an.

La survie attendue au temps t est calculée en faisant la moyenne des survies individuelles, $S_{ai}(t)$, des patients encore à risque au temps t :

$$S_a(t) = \frac{\sum_{i=1}^{n} Y_i(t) S_{ai}(t)}{\sum_{i=1}^{n} Y_i(t)}$$

avec $Y_i(t)$ étant l'indicatrice représentant le nombre de patient à risque au temps t.
Les sujets décédés avant le temps t ne contribuent plus au calcul de la survie attendue au temps t. Il ne s'agit donc pas de la survie attendue du collectif initial, la survie attendue est alors modifiée.
Cette méthode est toujours utilisée de nos jours dans des études nationales ou internationales - Eurocare V [De Angelis, 2014] - Seer [Howlader, 2014].

Cependant, même si le changement de structure de la population est considéré, particulièrement pour la méthode d'Ederer II, le fait qu'un patient qui n'est plus à risque au temps t peut être hypothétiquement encore vivant de son cancer au temps t n'est pas pris en compte, de la même façon que pour la survie spécifique ; Si le temps de survie dû au cancer et le temps de survie dû aux autres causes sont dépendants, un certain nombre de patients est retiré des patients à risque de manière non-aléatoire du fait des décès autres-causes ; ces derniers induisent donc une censure informative. Aucune des méthodes de survie relative ne prenant en compte ce type de censure, leur utilisation peut fournir des estimations biaisées de la survie nette.

II.1.2.b. Méthodes paramétriques du taux en excès

Les méthodes paramétriques du taux en excès raisonnent sur le taux de mortalité instantané. Le taux de mortalité observé au temps t s'exprime comme la somme de deux composantes : le taux de mortalité en excès au temps t, $\lambda_c(t)$, et le taux de mortalité autres causes au temps t, $\lambda_{\bar{c}}(t)$, tel que :

$$\lambda_o(t,x,z) = \lambda_c(t,z) + \lambda_{\bar{c}}(t,x)$$

avec x représentant les covariables démographiques et z représentant les covariables pronostiques.

Comme énoncé précédemment, les taux de mortalité autres-causes $\lambda_{\bar{c}}(t)$ sont approchés par les taux de mortalité attendus, $\lambda_a(t)$, d'une population présentant les mêmes caractéristiques démographiques mais qui n'est pas atteint du cancer en question. Le taux de mortalité dû au cancer est appelé taux de mortalité en excès car cela correspond à l'excès de mortalité que l'on veut observer dans la population malade comparé à une population saine présentant les mêmes caractéristiques démographiques. La population malade est inclue dans la population générale mais sa proportion étant très faible, les résultats ne sont pas influencés par ce fait [Ederer, 1961].

Modèle univarié

Dans le modèle univarié, aucune variable autre que le temps n'est incluse dans le modèle. Il s'agit de donner une forme paramétrique au taux de base. La stratégie proposée par Remontet [Remontet, 2007] repose sur le modèle suivant :

$$\log[\lambda_c(t)] = f_0(t)$$

Afin d'avoir une forme correspondant au mieux aux données, f_0 est choisie parmi plusieurs fonctions candidates plus ou moins souples à l'aide du critère d'Akaike [Akaike, 1973], telles qu'une fonction constante, une fonction linéaire, une fonction quadratique, une fonction cubique, un spline cubique de régression avec un nœud, un spline cubique de régression avec

plusieurs nœuds. Un spline cubique de régression est une fonction cubique par morceaux dont les deux premières dérivées sont continues aux nœuds. Cette fonction est reconnue pour sa flexibilité. Elle s'exprime de la manière suivante :

$$f(x) = \sum_{j=1}^{3} \beta_j x_j + \sum_{l=1}^{k} \beta_l (x - t_l)_+^3$$

où t_l, $l = \{1, \ldots, k\}$ sont les k nœuds.

Cette stratégie a été utilisée, il y a quelques années, dans les études menées par les registres du réseau Francim [Bossard, 2007].

Modèle multivarié

Dans le modèle multivarié, différentes covariables peuvent être incluses dans le modèle. Une forme paramétrique est donnée au taux de base mais également à l'effet des covariables. La stratégie proposée par Remontet [Remontet, 2007] repose sur le modèle dont l'écriture générale est la suivante :

$$\log[\lambda_c(t,z)] = f_0(t) + \beta(t) \times z + g(z)$$

De la même manière que le modèle univarié, le modèle le plus adéquat est choisi à l'aide du critère d'Akaike [Akaike, 1973] parmi plusieurs combinaison des fonctions f_0 , β et g.

Un modèle paramétrique prend en compte la censure informative en ajustant le modèle sur les covariables agissant à la fois sur le taux de mortalité en excès et sur le taux de mortalité dû aux autres causes, c'est-à-dire les variables démographiques (le modèle univarié ne prend pas en compte la censure informative ; les deux modèles sont équivalents uniquement si aucune covariable démographique n'agit sur le taux de mortalité en excès, ce qui est très rare). Cependant, pour modéliser au mieux le taux de mortalité en excès, une stratégie doit être élaborée concernant les différents points suivants : (i) Quelles covariables doivent être incluses dans le modèle ? (ii) Quelle forme donner au taux de mortalité de base ? (iii)

Comment modéliser l'effet des covariables sur la variable de sortie ? Constant au cours du temps ou dépendant du temps ? Si dépendant du temps, doit-on utiliser une fonction linéaire au cours du temps ou une fonction plus flexible telle qu'un spline ? Si la covariable est continue, son effet est-il linéaire ?, etc…. Les décisions prises sont importantes car un modèle inapproprié peut conduire à des conclusions erronées. Donc, si l'objectif de l'analyse de survie est uniquement une estimation ponctuelle, l'utilisation d'un modèle non-paramétrique permettant de prendre en compte la censure informative due à la dépendance entre les deux temps de survie aurait pour avantage important d'échapper à une stratégie de construction de modèle souvent délicate à mettre en place.

II.1.2.c. Méthode non paramétrique de Pohar-Perme

Un nouvel estimateur reposant sur la pondération par l'inverse de la probabilité de censure [Robins, 1993][Satten, 2001], écrit dans le cadre des processus de comptage, a été proposé récemment [Pohar-Perme, 2012] afin de prendre en compte la censure informative causée par l'influence d'une même covariable sur le temps de mortalité dû aux autres-causes et sur le temps de mortalité dû au cancer.

Intuitivement, en se positionnant dans le cadre de la survie spécifique, le nombre de patients à risque de décéder de leur cancer au temps *t* et le nombre de décès dû au cancer au temps *t* sont sous-estimés du fait que l'on ne suppose pas que les patients décédés des autres-causes puissent décéder du cancer dans un monde hypothétique.

Pour remédier à ce problème, Pohar & al propose de réajuster le nombre de décès dû au cancer au temps *t* ainsi que le nombre d'individus à risque au temps *t* en pondérant par la survie attendue afin de prendre en compte la proportion de personnes décédées des autres-causes dans la population à risque de décéder du cancer.

Dans le cadre de la cause de décès connue, l'estimateur de Nelson-Aalen pour le taux de mortalité cumulé en excès au temps *t* s'écrit de la manière suivante :

$$\hat{\Lambda}_c(t) = \int_0^t \frac{dN_c(u)}{Y(u)}$$

avec $N_c(t)$ le nombre de décès dû au cancer au temps t et $Y(t)$ le nombre de personne à risque au temps t.

En approchant le taux de mortalité autres-causes de la population par le taux de mortalité attendu d'une population saine présentant les mêmes caractéristiques démographiques, l'estimateur de Nelson-Aalen pondéré est alors le suivant :

$$\hat{\Lambda}_c(t) = \int_0^t \frac{dN_c^w(u)}{Y^w(u)}$$

avec $N_c^w(u) = \sum_i N_{ci}^w(u) = \sum_i \frac{N_{ci}(u)}{S_{ai}(u)}$ et $Y_c^w(u) = \sum_i Y_{ci}^w(u) = \sum_i \frac{Y_{ci}(u)}{S_{ai}(u)}$, $N_{ci}(t)$ et $Y_{ci}(t)$ étant respectivement l'indicatrice individuelle d'évènement au temps t et l'indicatrice individuelle d'être à risque dans la population au temps t.

Pour estimer la survie nette sans avoir la cause de décès (cadre de la cause de décès inconnue), le nouvel estimateur propose de pondérer l'expression du taux de mortalité cumulé en excès obtenu par la méthode d'Ederer II de la même manière que précédemment :

Ainsi, à partir de l'expression du taux de mortalité en excès de la méthode d'Ederer II [Pohar-Perme, 2012] :

$$\hat{\Lambda}_c(t) = \int_0^t \frac{dN(u)}{Y(u)} - \int_0^t \frac{\sum_{i=1}^n Y_i(u) d\Lambda_{ai}(u)}{Y(u)}$$

L'expression du taux de mortalité cumulé en excès du nouvel estimateur est la suivante :

$$\hat{\Lambda}_c(t) = \int_0^t \frac{dN^w(u)}{Y^w(u)} - \int_0^t \frac{\sum_{i=1}^n Y_i^w(u) d\Lambda_{ai}(u)}{Y^w(u)}$$

Cette méthode est non-paramétrique et ne nécessite aucune modélisation.

II.2. Evaluation des performances des différents estimateurs de la survie nette existants

II.2.1. Comment juger des performances d'un estimateur ?

L'évaluation des performances d'un estimateur peut être réalisée à l'aide d'études théoriques, ce qui a déjà été fait dans le cadre de la survie nette par Pohar & al [Pohar-Perme, 2012] ou à l'aide d'études de simulations en utilisant des critères de performances jugés pertinents tels que le biais, l'erreur quadratique moyenne ou la racine carrée de l'erreur quadratique moyenne, que l'on nommera dans la suite RMSE, ou encore le taux de couverture des intervalles de confiance. C'est sur ce deuxième type d'étude que va porter la suite de ce chapitre.

Il est important de savoir que les études empiriques ne permettent pas d'évaluer les performances d'un estimateur. Elles peuvent seulement mesurer l'impact de l'utilisation d'une méthode biaisée sur données réelles.

Le biais

Théoriquement, en supposant $\hat{\theta}_n$ un estimateur de θ, le biais entre $\hat{\theta}_n$ et θ est une fonction B_n telle que :

$$B_n : \Theta \times \Theta \to R$$
$$\left(\theta, \hat{\theta}_n\right) \mapsto E\left(\hat{\theta}_n\right) - \theta$$

Si $B_n = 0 \; \forall \left(\theta, \theta_n\right) \in \Theta \times \Theta$, on dit que $\hat{\theta}_n$ est un estimateur sans biais de θ.

Dans une étude de simulation, le biais d'un estimateur peut être estimé comme la différence entre la valeur théorique et la moyenne de cet estimateur, qui correspond dans notre cas, à la moyenne des valeurs estimées :

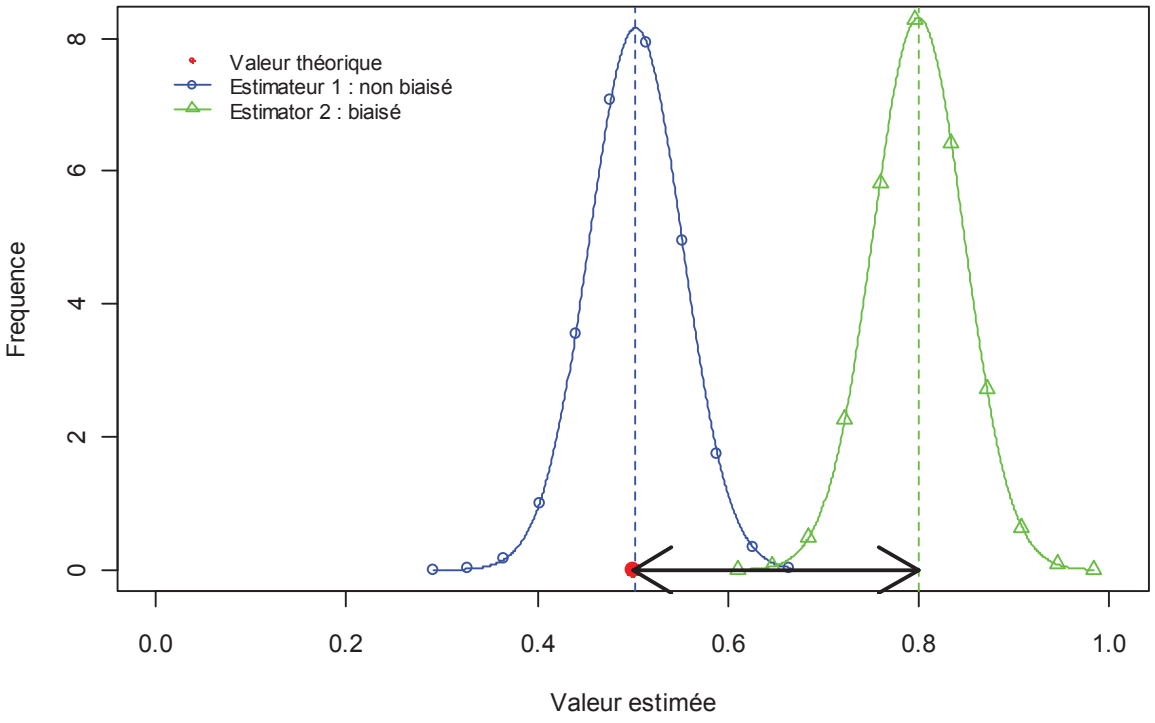
$$Biais = \frac{1}{M}\sum_{i=1}^{M}\hat{s}_i - s^*$$

avec $\hat{s}_i$ représentant la valeur estimée pour le $i^{ème}$ jeu de données simulés et s^* représentant la valeur théorique que l'on cherche à estimer.

Prenons un exemple afin d'avoir une illustration concrète :

Voici le graphe en figure II.1 de la distribution des valeurs estimées pour 1000 jeux de données avec l'estimateur 1 (ronds bleus) et l'estimateur 2 (triangles verts), la valeur théorique, égale à 0.5, étant représentée en point rouge.

Figure II.1. Distribution des valeurs de survie estimées pour 1000 jeux de données simulés



Nous pouvons remarquer que l'estimateur 1 n'est pas biaisé ; la différence entre la valeur théorique et la moyenne des valeurs empiriques est nulle. En revanche, nous ne pouvons pas dire la même chose du second estimateur. L'estimateur 2 est biaisé, l'amplitude de son biais étant représenté par la flèche. Dans ce cas, du fait qu'ils ont la même variance, il semble plus pertinent d'utiliser l'estimateur 1 que l'estimateur 2.

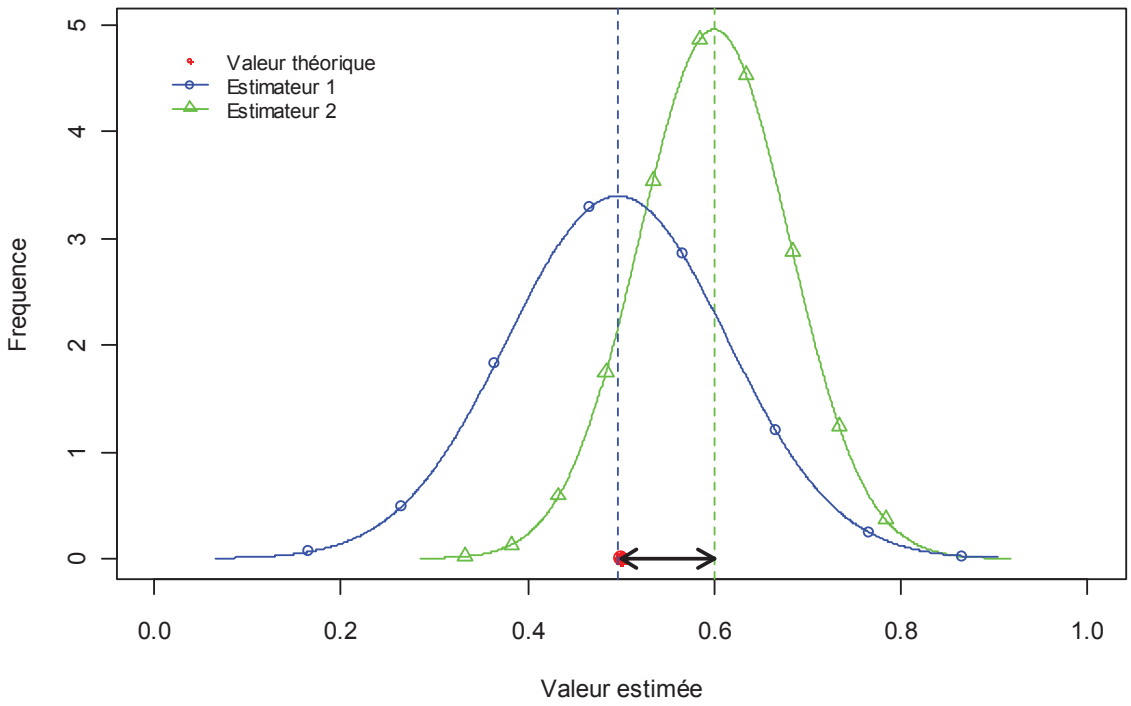
L'erreur quadratique moyenne

Cependant, être en possession de plusieurs estimateurs de même variance est une situation rare. Il arrive le plus souvent de se trouver dans le cas où nous devons choisir entre un estimateur très peu biaisé mais possédant une assez large variance et un estimateur biaisé mais n'ayant pas une grande variance (figure II.2). Comment faire pour savoir lequel des deux est le plus pertinent ?

Une solution est d'utiliser la RMSE. Théoriquement, en supposant $\hat{\theta}_n$ un estimateur de θ, la RMSE est la fonction $RMSE_n$ telle que :

$$RMSE_n : \Theta \times \Theta \to R$$
$$(\theta, \hat{\theta}_n) \mapsto E\left((\hat{\theta}_n - \theta)^2\right)$$

Figure II.2. Distribution des valeurs de survie estimées pour 1000 jeux de données simulés



La RMSE représente, sur données simulées, la moyenne du carré de la différence entre les valeurs estimées et la valeur théorique :

$$RMSE = \frac{1}{M}\sum_{i=1}^{M}\left(\left(\hat{s}_i - s^*\right)^2\right)$$

$$RMSE = \left[\frac{1}{M}\sum_{i=1}^{M}\hat{s}_i - s^*\right]^2 + \mathrm{var}\left(\hat{s}_i\right)$$

Elle résume l'information en combinant l'information du biais et de la variabilité.

Dans l'exemple de la figure II.2, la valeur de la RMSE de l'estimateur 2 est de 0.014 alors que celle de l'estimateur 1 est de 0.016. Dans ce cas, même si l'estimateur 2 est biaisé, il est retenu comme étant l'estimateur le plus performant selon le critère du RMSE

Le taux de couverture

Le taux de couverture représente le pourcentage d'intervalle de confiance qui contient la valeur théorique. Un bon estimateur doit avoir un taux de couverture proche de la valeur nominal des 95%. A noter qu'un intervalle de confiance étroit observé sur données réelles n'est aucunement un critère de performance : en effet, il pourrait ne pas inclure la valeur théorique avec une probabilité de 95%, première propriété essentielle que l'on requière d'un intervalle de confiance.

II.2.2. Quantités théoriques en jeu

II.2.2.a. Quantité théorique d'intérêt dans un monde hypothétique où la seule cause de décès serait le cancer

La survie nette, notre quantité d'intérêt, qui représente la survie que l'on observerait dans la situation hypothétique où la seule cause de mortalité possible serait le cancer étudié, est associée au taux de mortalité individuel dû au cancer de la population, $\lambda_{c,i}$, qui s'exprime de la façon suivante :

$$\lambda_{c,i}(t) = \lim_{dt \to 0} \frac{P\left(t < T_{c,i} \leq t + dt \middle| T_{c,i} > t\right)}{dt}$$

avec $T_{c,i}$ représentant le temps de décès dû au cancer pour le patient i.

$\lambda_{c,i}(t)dt$ correspond, pour le patient i, à la probabilité de décéder du cancer entre t et $t + dt$ conditionnellement au fait d'être encore vivant de ce cancer en t.

A l'aide de la relation usuelle entre la survie et le taux de mortalité, la survie nette individuelle s'exprime de la manière suivante :

$$S_{c,i}(t) = \exp\left(- \int_0^t \lambda_{c,i}(u) du \right)$$

Or, le taux de mortalité individuel n'étant pas identique d'un individu à l'autre, le taux de mortalité de la population s'exprime alors de la façon suivante :

$$\lambda_c(t) = \frac{\sum_i S_{c,i}(t) \lambda_{c,i}(t)}{\sum_i S_{c,i}(t)} \qquad \text{(II.1)}$$

la survie nette associée étant :

$$S_c(t) = \exp\left(- \int_0^t \lambda_c(u) du \right).$$

II.2.2.b. Quantités théoriques rencontrées dans le « monde réel » dans lequel il existe des risques compétitifs

Cause de décès connue

En présence de risques compétitifs et lorsque la cause de décès est connue, le taux de mortalité individuel spécifique au cancer représente la quantité suivante :

$$\lambda^*_{c,i}(t) = \lim_{dt \to 0} \frac{P\left(t < T_{c,i} \leq t + dt \middle| T_i > t\right)}{dt} \qquad \text{(II.2)}$$

avec $T_{c,i}$ représentant le temps de décès dû au cancer pour le patient i et T_i représentant le minimum entre $T_{c,i}$ et $T_{\bar{c},i}$, temps de décès dû aux autres-causes, pour le patient i.

$\lambda_{c,i}{}^*(t)dt$ représente, pour le patient i, la probabilité de décéder du cancer dans l'intervalle de temps $t + dt$ conditionnellement au fait d'être encore vivant en t.

Le taux de mortalité de la population spécifique au cancer, $\lambda_c{}^*$, s'exprime de la manière suivante :

$$\lambda^*_c(t) = \frac{\sum_i S_{o,i}(t)\lambda_{c,i}(t)}{\sum_i S_{o,i}(t)}$$

$$\lambda^*_c(t) = \frac{\sum_i S_{o,i}(t)\lambda_{o,i}(t)}{\sum_i S_{o,i}(t)} - \frac{\sum_i S_{o,i}(t)\lambda_{\bar{c},i}(t)}{\sum_i S_{o,i}(t)}$$

avec $S_{o,i}(t)$ la survie observée du patient i au temps t, $\lambda_{o,i}(t)$ le taux de mortalité observé du patient i au temps t et $\lambda_{\bar{c}}(t)$ le taux de mortalité autres-causes du patient i au temps t.

Se situant dans le cadre des risques concurrents, la survie spécifique, associée au taux de mortalité spécifique, peut également être vue comme la survie nette observable. La survie

nette observable est associée à la probabilité de survenue de décès dus au cancer en présence de la survenue de décès autres-causes, appelée probabilité brute de décès dus au cancer.
Nous pouvons ajouter que dépendant de la survie observée et donc des autres-causes, cette quantité ne permet pas de comparaison entre pays ou période de diagnostic.

Cause de décès inconnue

Une autre quantité que l'on peut exprimer dans le monde réel en présence de risques compétitifs est la survie relative, c'est-à-dire, le rapport entre la survie observée et la survie attendue ; il s'écrit de la manière suivante :

$$S_r(t) = \frac{\sum_{i=1}^{n} S_{o,i}(t)}{\sum_{i=1}^{n} S_{a,i}(t)}$$

$$S_r(t) = \frac{\sum_{i=1}^{n} \exp\left(-\int_0^t \lambda_{o,i}(u)du\right)}{\sum_{i=1}^{n} \exp\left(-\int_0^t \lambda_{a,i}(u)du\right)}$$

$$S_r(t) = \exp\left(-\int_0^t \lambda_c^{**}(u)du\right)$$

La quantité liée à la survie relative est donc :

$$\lambda_c^{**}(t) = \frac{\sum_{i=1}^{n} S_{o,i}(t)\lambda_{o,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)} - \frac{\sum_{i=1}^{n} S_{a,i}(t)\lambda_{a,i}(t)}{\sum_{i=1}^{n} S_{a,i}(t)} \qquad \text{(II.3)}$$

Contrairement à $\lambda_c(t)dt$ et $\lambda_c^{*}(t)dt$, $\lambda_c^{**}(t)dt$ ne représente pas une probabilité car $\lambda_c^{**}(t)$ peut être négatif.

L'interprétation de la survie relative est la suivante : une survie relative à 5 ans de suivi égale à 0.7 signifie que 5 ans après le diagnostic, la survie des patients suivis est égale à 0.7 fois celle de personnes présentant les mêmes caractéristiques démographiques dans la population

générale. Le fait qu'elle dépende des autres-causes de mortalité ne permet aucune comparaison entre pays ou période de diagnostic. Nous pouvons ajouter également que cette quantité pouvant être supérieure à 1, il ne s'agit pas d'une probabilité de survie.

II.2.3. Relation entre les quantités théoriques en jeu selon différents cas de figure

Dans cette partie, nous allons voir sous quelles conditions les différentes quantités théoriques correspondent à la survie nette.

Voici, en table II.1, un récapitulatif des différentes quantités en jeu :

Table II.1. Quantités théoriques en jeu

Quantités théoriques	« Taux » associés
Survie Nette	$\lambda_c(t) = \frac{\sum_i S_{c,i}(t)\lambda_{c,i}(t)}{\sum_i S_{c,i}(t)}$
Survie Nette Observable (Survie spécifique)	$\lambda_c^*(t) = \frac{\sum_i S_{o,i}(t)\lambda_{o,i}(t)}{\sum_i S_{o,i}(t)} - \frac{\sum_i S_{o,i}(t)\lambda_{\bar{c},i}(t)}{\sum_i S_{o,i}(t)} = \frac{\sum_i S_{o,i}(t)\lambda_{c,i}(t)}{\sum_i S_{o,i}(t)}$
Survie Relative	$\lambda_c^{**}(t) = \frac{\sum_i S_{o,i}(t)\lambda_{o,i}(t)}{\sum_i S_{o,i}(t)} - \frac{\sum_i S_{a,i}(t)\lambda_{a,i}(t)}{\sum_i S_{a,i}(t)}$

II.2.3.a. Le taux de mortalité en excès est homogène dans la population

En reprenant les expressions des taux de mortalité en excès des différentes méthodes de la table II.1, nous pouvons voir que lorsque le taux de mortalité est homogène dans la population, c'est-à-dire lorsque $\lambda_{c,i}(t) = \lambda_c(t)$ $\forall$i, les quantités définies par les différentes méthodes correspondent à la survie nette. En effet :

Survie Nette

Le taux de mortalité en excès associé à la survie nette est alors le suivant :

$$\lambda_c^{SN}(t) = \frac{\sum_i S_{c,i}(t)\lambda_{c,i}(t)}{\sum_i S_{c,i}(t)}$$

$$\lambda_c^{SN}(t) = \lambda_c(t)$$

Survie Nette Observable

Le taux de mortalité en excès associé à la survie nette observable est alors le suivant :

$$\lambda_c^{SNO}(t) = \lambda_o(t) - \frac{\sum_i S_{o,i}(t)\lambda_{\bar{c},i}(t)}{\sum_i S_{o,i}(t)}$$

$$\lambda_c^{SNO}(t) = \frac{\sum_i S_{o,i}(t)\lambda_{o,i}(t)}{\sum_i S_{o,i}(t)} - \frac{\sum_i S_{o,i}(t)\lambda_{\bar{c},i}(t)}{\sum_i S_{o,i}(t)}$$

$$\lambda_c^{SNO}(t) = \frac{\sum_i S_{o,i}(t)\lambda_{c,i}(t)}{\sum_i S_{o,i}(t)}$$

$$\lambda_c^{SNO}(t) = \frac{\lambda_c(t)\sum_i S_{o,i}(t)}{\sum_i S_{o,i}(t)}$$

$$\lambda_c^{SNO}(t) = \lambda_c(t)$$

Survie Relative

Lorsque le taux de mortalité en excès est homogène pour la population considérée, le taux de mortalité en excès associé à la survie relative est alors le suivant :

$$\lambda_c^{SR}(t) = \lambda_o(t) - \frac{\sum_i S_{a,i}(t)\lambda_{a,i}(t)}{\sum_i S_{a,i}(t)}$$
$$\lambda_c^{SR}(t) = \lambda_c(t) + \lambda_a(t) - \lambda_a(t)$$
$$\lambda_c^{SR}(t) = \lambda_c(t)$$

En conclusion, dans ce cas, le taux de mortalité en excès associé à chaque quantité correspond au taux de mortalité associé à la survie nette. La survie nette observable et la survie relative correspondent donc à la survie nette.

II.2.3.b. Le taux de mortalité en excès est hétérogène dans la population

Lorsque le taux de mortalité en excès est hétérogène dans la population, les différentes quantités du tableau II.1 ne correspondent pas à la même chose. Comment évolue le taux de mortalité en excès associé à la survie nette observable et à la survie relative par rapport au taux de mortalité en excès associé à la survie nette, quantité théorique que l'on cherche à estimer ?

Comparaison survie nette et survie nette observable

Dans le cadre de la survie nette observable, les taux de mortalité individuels dus au cancer sont pondérés par la probabilité que le patient *i* soit encore à risque au temps *t* parmi les patients encore à risque au temps *t* :

$$\frac{S_{o,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)}$$

Dans le cadre de la survie nette, les taux de mortalité individuels dus au cancer sont pondérés par la probabilité que le patient *i* survive de son cancer au temps *t* parmi les patients survivants de leur cancer au temps *t* :

$$\frac{S_{c,i}(t)}{\sum_{i=1}^{n} S_{c,i}(t)}$$

En supposant que la covariable *âge* ait un effet sur le taux de mortalité en excès, la répartition des âges ne sera pas la même selon que l'on se situe dans le monde réel ou dans le monde hypothétique où le cancer serait la seule cause de mortalité.

De manière générale, les personnes âgées sont les personnes les plus à risque de décéder des autres-causes mais également du cancer, on a alors l'inégalité suivante :

$$\frac{S_{o,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)} < \frac{S_{c,i}(t)}{\sum_{i=1}^{n} S_{c,i}(t)}$$

$$\frac{S_{o,i}(t)\lambda_{c,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)} < \frac{S_{c,i}(t)\lambda_{c,i}(t)}{\sum_{i=1}^{n} S_{c,i}(t)}$$

Cela signifie que le poids d'une personne âgée *i* est plus faible pour la survie nette observable que pour la survie nette.

Au final,

$$\lambda_c^{SNO}(t) < \lambda_c^{SN}(t)$$

Cette inégalité s'inverse pour les patients jeunes.

Voici une illustration sur deux patients :

	$S_{c,i}$	$S_{a,i}$	$S_{o,i}$
Patient Jeune	0.95	0.95	0.9025
Patient Vieux	0.5	0.5	0.25

Pour le patient âgé,

$$\frac{S_{o,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)} = \frac{0.25}{0.9025 + 0.25} = 0.21$$

$$\frac{S_{c,i}(t)}{\sum_{i=1}^{n} S_{c,i}(t)} = \frac{0.5}{0.95 + 0.5} = 0.34$$

On a bien :

$$\frac{S_{o,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)} < \frac{S_{c,i}(t)}{\sum_{i=1}^{n} S_{c,i}(t)}$$

En revanche, pour le patient jeune,

$$\frac{S_{o,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)} = \frac{0.9025}{0.9025 + 0.25} = 0.78$$

$$\frac{S_{c,i}(t)}{\sum_{i=1}^{n} S_{c,i}(t)} = \frac{0.95}{0.95 + 0.5} = 0.66$$

Dans ce cas :

$$\frac{S_{o,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)} > \frac{S_{c,i}(t)}{\sum_{i=1}^{n} S_{c,i}(t)}$$

Comparaison survie nette observable et survie relative

Dans le cadre de la survie nette observable, en supposant que les taux de mortalité autres-causes $\lambda_{\bar{c}}$ sont bien reflétés par le taux de mortalité de la population générale λ_a, les taux de

mortalité attendus individuels sont pondérés par la probabilité que le patient *i* soit encore à risque au temps *t* parmi les patients encore à risque au temps *t* :

$$\frac{S_{o,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)}$$

Dans le cadre de la survie relative, les taux de mortalité attendus individuels sont pondérés par la probabilité de survie attendue du patient *i* au temps *t* parmi les patients survivants aux autres causes au temps *t* :

$$\frac{S_{a,i}(t)}{\sum_{i=1}^{N} S_{a,i}(t)}$$

La covariable *âge* ayant un effet sur le taux de mortalité attendu, la répartition des âges n'est pas la même selon que l'on se situe dans le monde réel ou dans le monde hypothétique où les patients ne décéderaient que des autres-causes.

En prenant l'exemple des personnes âgées, dans le cas où celles-ci seraient les personnes les plus à risque de décéder des autres-causes mais également du cancer, on a l'inégalité suivante :

$$\frac{S_{o,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)} < \frac{S_{a,i}(t)}{\sum_{i=1}^{n} S_{a,i}(t)}$$

$$\frac{S_{o,i}(t)\lambda_{a,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)} < \frac{S_{a,i}(t)\lambda_{a,i}(t)}{\sum_{i=1}^{n} S_{a,i}(t)}$$

$$-\frac{S_{o,i}(t)\lambda_{a,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)} > -\frac{S_{a,i}(t)\lambda_{a,i}(t)}{\sum_{i=1}^{n} S_{a,i}(t)}$$

$$\lambda_o(t) - \frac{\sum_i S_{o,i}(t)\lambda_{a,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)} > \lambda_o(t) - \frac{\sum_i S_{a,i}(t)\lambda_{a,i}(t)}{\sum_{i=1}^{n} S_{a,i}(t)}$$

Alors,

$$\lambda_c^{SNO}(t) > \lambda_c^{SR}(t)$$

Voici une illustration sur deux patients :

	$S_{c,i}$	$S_{a,i}$	$S_{o,i}$
Patient Jeune	0.95	0.95	0.9025
Patient Vieux	0.5	0.5	0.25

Pour le patient âgé,

$$\frac{S_{o,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)} = \frac{0.25}{0.9025 + 0.25} = 0.21$$

$$\frac{S_{a,i}(t)}{\sum_{i=1}^{n} S_{a,i}(t)} = \frac{0.5}{0.95 + 0.5} = 0.34$$

On a bien :

$$\frac{S_{o,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)} < \frac{S_{a,i}(t)}{\sum_{i=1}^{n} S_{a,i}(t)}$$

En revanche, pour le patient jeune,

$$\frac{S_{o,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)} = \frac{0.9025}{0.9025 + 0.25} = 0.78$$
$$\frac{S_{a,i}(t)}{\sum_{i=1}^{n} S_{a,i}(t)} = \frac{0.95}{0.95 + 0.5} = 0.66$$

Dans ce cas :

$$\frac{S_{o,i}(t)}{\sum_{i=1}^{n} S_{o,i}(t)} > \frac{S_{a,i}(t)}{\sum_{i=1}^{n} S_{a,i}(t)}$$

Comparaison survie nette et survie relative

A l'aide des deux sous-parties ci-dessus,

Pour une population âgée : $\lambda_c^{SN}(t) > \lambda_c^{SNO}(t) > \lambda_c^{SR}(t) \quad \Rightarrow \quad S_c^{SN}(t) < S_c^{SNO}(t) < S_c^{SR}(t)$

Pour une population jeune : $\lambda_c^{SN}(t) < \lambda_c^{SNO}(t) < \lambda_c^{SR}(t) \quad \Rightarrow \quad S_c^{SN}(t) > S_c^{SNO}(t) > S_c^{SR}(t)$

La survie relative et la survie nette observable ne correspondent pas à la survie nette dans ce cas. La survie relative semble être plus éloignée que la survie nette observable.

II.2.4. Vers quelles quantités théoriques convergent les estimateurs ?

II.2.4.a. Méthodes d'Ederer I et de Hakulinen

Pohar-Perme a montré [Pohar-Perme, 2012] que le taux de mortalité estimé par la méthode d'Ederer I et par la méthode de Hakulinen était un estimateur consistent de la quantité :

$$\lambda_c^{**}(t) = \frac{\sum_i S_{o,i}(t)\lambda_{o,i}(t)}{\sum_i S_{o,i}(t)} - \frac{\sum_i S_{a,i}(t)\lambda_{a,i}(t)}{\sum_i S_{a,i}(t)}$$

Les méthodes d'Ederer I et de Hakulinen sont donc associée à la survie relative.

D'après la partie II.2.3, les méthodes d'Ederer I et de Hakulinen n'estiment pas la survie nette en cas d'hétérogénéité. Ces méthodes estiment une quantité qui dépend du taux de mortalité attendu de la population et donc qui ne permet pas de comparaison entre pays ou périodes de diagnostic.

II.2.4.b. Méthodes d'Ederer II

Pohar & al a montré [Pohar-Perme, 2012] que le taux de mortalité estimé par la méthode d'Ederer II était un estimateur consistent de la quantité :

$$\lambda_c^{*}(t) = \frac{\sum_i S_{o,i}(t)\lambda_{c,i}(t)}{\sum_i S_{o,i}(t)}$$

Lorsque les taux de mortalité autres-causes sont approximés par les taux de mortalité de la population générale, la méthode d'Ederer II et la survie spécifique sont équivalentes.
D'après la partie II.2.3, la méthode d'Ederer II est donc associée à la survie nette observable et elle n'estime donc pas la survie nette en cas d'hétérogénéité. Les méthodes associées à la survie nette observable estiment une quantité qui dépend de la survie observée et donc des autres-causes, ce qui ne permet pas de comparaison entre pays ou période de diagnostic.

II.2.4.c. Méthode de Pohar

Pohar & al a montré [Pohar-Perme, 2012] que le taux de mortalité estimé par le nouvel estimateur était un estimateur consistent de la quantité :

$$\lambda_c(t) = \frac{\sum_{i=1}^{n} S_{c,i}(t)\lambda_{c,i}(t)}{\sum_{i=1}^{n} S_{c,i}(t)}$$

L'estimateur de Pohar-Perme est un estimateur non biaisé de la survie nette en cas d'hétérogénéité, ce qui n'est pas le cas pour les autres méthodes.

II.2.5. Comparaison des méthodes d'estimation de la survie nette à l'aide d'une étude de simulation

L'article présenté dans ce paragraphe porte sur l'étude de la performance des méthodes d'estimation de la survie nette sur données simulées. Le but est de pouvoir rendre compte des erreurs commises sur l'estimation de la survie nette en utilisant des méthodes biaisées. En effet, comme nous l'avons vu dans le paragraphe précédent, les méthodes de survie relative et celles représentant la survie nette observable estiment une quantité différente de la survie nette dans le cadre de la présence d'une censure informative. Cet article vient alors en complément de l'article de Pohar & al [Pohar-Perme, 2012] présentant l'étude théorique. Depuis le début, un amalgame s'est installé entre survie nette et survie relative. L'un des messages les plus importants de ces études est qu'il ne faut plus parler de survie relative en terme de concept car ce n'est pas ce que l'on recherche à estimer et que les méthodes de survie relative ne doivent plus être utilisées pour estimer la survie nette.
Cet article a été publié dans la revue *Statistics in Medicine*.

Research Article

Statistics in Medicine

Received 8 April 2011, Accepted 20 October 2011 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4464

Estimating net survival: the importance of allowing for informative censoring

Coraline Danieli,[a,b,c,d*†] Laurent Remontet,[a,b,c,d] Nadine Bossard,[a,b,c,d] Laurent Roche[a,b,c,d] and Aurélien Belot[a,b,c,d,e]

Net survival, the one that would be observed if cancer were the only cause of death, is the most appropriate indicator to compare cancer mortality between areas or countries. Several parametric and non-parametric methods have been developed to estimate net survival, particularly when the cause of death is unknown. These methods are based either on the relative survival ratio or on the additive excess hazard model, the latter using the general population mortality hazard to estimate the excess mortality hazard (the hazard related to net survival). The present work used simulations to compare estimator abilities to estimate net survival in different settings such as the presence/absence of an age effect on the excess mortality hazard or on the potential time of follow-up, knowing that this covariate has an effect on the general population mortality hazard too. It showed that when age affected the excess mortality hazard, most estimators, including specific survival, were biased. Only two estimators were appropriate to estimate net survival. The first is based on a multivariable excess hazard model that includes age as covariate. The second is non-parametric and is based on the inverse probability weighting. These estimators take differently into account the informative censoring induced by the expected mortality process. The former offers great flexibility whereas the latter requires neither the assumption of a specific distribution nor a model-building strategy. Because of its simplicity and availability in commonly used software, the non-parametric estimator should be considered by cancer registries for population-based studies. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: net survival; relative survival; excess hazard model; informative censoring

1. Introduction

Survival is an important indicator of the efficiency of a health care system in disease management, especially cancer care. Together with other indicators such as mortality, incidence and prevalence, it provides important information on the burden of cancer in a given population. Survival is also used in clinical trials to evaluate cancer treatments and in epidemiology to identify the determinants of survival differences and trends.

In cancer research, the overall survival probability considers death as an event whatever its cause; it is thus influenced by the mortality hazard because of other causes of death than the cancer under study, especially when the studied population includes a high proportion of elderly people. In exploring the impact of a specific type of cancer on a population, it is rather interesting to estimate the net survival probability defined as survival if cancer were the only cause of death. This net survival probability should be used for comparisons of cancer impact between countries or time periods because it is not influenced by death because of other causes.

The estimation of net survival, an unobservable indicator, is rather difficult, even when the cause of death is known. In this cause-specific setting, the classical approach considers the other causes of death as censoring events and uses Kaplan–Meier method. The value of this estimator is limited because the

[a]*Hospices Civils de Lyon, Service de Biostatistique, F-69003, Lyon, France*
[b]*Université de Lyon, F-69000, Lyon, France*
[c]*Université Lyon 1, F-69100, Villeurbanne, France*
[d]*CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biotatistique-Santé, F-69100, Villeurbanne, France*
[e]*Institut de Veille Sanitaire, Département des maladies chroniques et traumatismes, F-94410, Saint-Maurice, France*
**Correspondence to: Coraline Danieli, Service de Biostatistique - Hospices Civils de Lyon, 165, chemin du Grand Revoyet, Bât. 4D, F-69495 Pierre-Bénite Cedex, France.*
†E-mail: coraline.danieli@chu-lyon.fr

causes of death of some patients may be missing or, when not, may be inaccurate, unreliable (especially in patients with multiple pathologies) [1, 2] or highly dependent on the local coding practices.

Furthermore, a frequent assumption in survival analyses is that the censoring process is independent from event occurrence [3, 4]. In estimating net survival in the cause-specific setting, this assumption may not be true. This occurs when one or several covariates influence simultaneously the two mortality hazard components: the cancer-specific mortality hazard and the other-causes mortality hazard (e.g., age). Because old people are likely to be more censored than young ones by death because of other causes, the censoring process becomes informative and leads to biased estimates of the net survival [5].

To avoid the need for the cause of death, two main approaches have been proposed. Both assume that the other-causes mortality hazard of each patient can be obtained from the life table of the general population; this hazard being that of an individual of same sex, age, *Département* of residence and other covariates present in the life table.

In the hazard modelling approach [6–9], the observed total mortality hazard is split into two components: the 'general population hazard' (i.e., the 'expected hazard') and the cancer-specific hazard; that is, the excess hazard because of cancer. A key assumption in this approach is that these two mortality processes should be independent (as in the cause-specific setting). When the excess hazard is assumed identical to all patients, this key assumption holds and one may use a non-parametric estimate of net survival as proposed by Andersen and Vaeth [8], which is close to the earlier proposal of Ederer and Heise [10]. However, when the excess hazard depends on one or several life-table variables, informative censoring may occur and an unbiased estimate of the net survival for the whole group requires taking into account this (these) covariate(s) in the model. This can be done by including this (these) covariate(s) in a multivariable excess hazard model in which the above key assumption becomes an assumption of independence conditionally on this (these) covariate(s). Many variants of the multivariable excess hazard model were developed in the last two decades. The baseline hazard was modelled with continuous and flexible parametric functions and the non-proportional effect of the covariates taken into account [11–15]. Semi-parametric excess hazard models were also proposed [16, 17].

The second approach proposed in absence of knowledge of the cause of death is the relative survival ratio [18, 19]. This ratio compares the observed survival to the survival that the patients would have experienced in the absence of cancer; it does not require the cause of death. This 'expected' survival is calculated in several slightly different ways, the common idea being the best possible match between an observed and a reference population. This is difficult to achieve because the patients with the lowest survival are withdrawn early from the sample and their relative survival ratio tends to equal that of the patients with the highest expected survival. The relative survival ratio is thus a biased estimator of the net survival.

Another problem arising in many population-based studies is the change over time of the distributions of the life-table variables among incident cancer patients. With a common closing date of follow-up, these changes imply that the potential follow-up times are heterogeneous and depend on the life-table variables. For example, if the number of old patients increases over time and that of young patients decreases, old patients will be, on average, censored earlier than young patients, inducing informative censoring (as the censoring times will depend on age).

The present study uses simulations to compare the performance of several estimators of the net survival probability, especially in case of informative censoring because of life-table variables (here age) and/or heterogeneous potential follow-up times. These estimators are described in Section 2. Different scenarios were explored to mimic real data, with or without informative censoring because of the effect of some covariates on both the excess mortality hazard and the general population mortality hazard. These scenarios allowed evaluating the statistical performance of each of the above-cited estimators.

Section 3 presents the simulation design of each scenario, the data generation process, the theoretical parameters and the criteria used to evaluate the performance of the estimators. The results are shown in Section 4 and discussed in Section 5.

2. Methods

2.1. Non-parametric estimators

$\hat{S}_e(t)$, the non-parametric estimators of net survival at time t, herein called ratio-estimators, are defined as

$$\hat{S}_e(t) = \frac{\hat{S}_o(t)}{S_p(t)} \quad (1)$$

where $\hat{S}_o(t)$ is an estimator of the crude survival of the group at time t and $S_P(t)$ the expected survival or population survival of the group at time t (i.e., the survival of the group if they were disease-free). The estimation of $\hat{S}_e(t)$ differs according to the way $S_P(t)$ is formulated. For lack of computer power, they were initially computed after splitting the follow-up time into short time intervals. In the present article, the calculation of the expected survival is done at each event (death or censorship).

2.1.1. Ederer I estimator. Ederer I estimator [18] estimates the population survival of a group of patients at time t as the average

$$S_P(t) = \sum_{i=1}^{n} \frac{S_{Pi}(t)}{n}$$

where $S_{Pi}(t)$ is the expected survival of a subject from the general population who has, at diagnosis, the same life-table characteristics as patient i.

2.1.2. Ederer II estimator. Ederer II estimator [10] estimates the population survival of a group of patients at time t by

$$S_P(t) = \frac{\sum_{i=1}^{n} Y_i(t).S_{Pi}(t)}{\sum_{i=1}^{n} Y_i(t)}.$$

where $Y_i(t) = 1$ if patient i is at risk of death at time t and 0 otherwise. This calculation of the population survival aimed initially at matching the age distribution of the disease-free group to that of the surviving patients [18, 19].

2.1.3. Hakulinen estimator. Hakulinen estimator [19] takes into account the potential follow-up time; that is, the maximum time during which a patient can be observed. The population survival is estimated by

$$S_p(t) = \frac{\sum_{i=1}^{n} C_i(t) \cdot S_{Pi}(t)}{\sum_{i=1}^{n} C_i(t)}$$

where $C_i(t) = 1$ if t is less (or equal to) the potential follow-up time and 0 otherwise. This method aims to take into account the informative censoring because of factors that influence simultaneously the survival and the potential follow-up time. In Equation (1), the method tends to bias the denominator in the same way as the numerator.

The two estimators that follow are nonparametric estimators but are not based on the ratio between $\hat{S}_o(t)$ and $S_P(t)$.

2.1.4. Pohar-Perme estimator. Pohar-Perme estimator is a weighted version of the Ederer II estimator [5]. The cumulative excess hazard of Ederer II estimator $\hat{\Lambda}_e(t)$ can be written, in a counting process approach, as the difference between the Nelson–Aalen estimate [3] and the cumulative population hazard of the patients still at risk at each death [8]

$$\hat{\Lambda}_e(t) = \int_0^t \frac{dN(u)}{Y(u)} - \int_0^t \frac{\sum_{i=1}^{n} Y_i(u) d\Lambda_{Pi}(u)}{Y(u)} \tag{2}$$

where $Y(u) = \sum_{i=1}^{n} Y_i(u)$ and $N(u) = \sum_{i=1}^{n} N_i(u)$, $N(u)$ being the number of deaths up to and including the time u.

To eliminate the bias present in the risk set $\{Y_i(t), i = 1...n\}$, each $Y_i(t)$ is increased by dividing it by $S_{Pi}(t)$: the lower is the probability of remaining at risk, the greater is the increase [5]. The counting

process $N_i(t)$ is modified in the same way. This procedure is based on the inverse probability weighting [20]. The Pohar-Perme estimator of the net survival is then defined by its cumulative excess hazard

$$\hat{\Lambda}_e(t) = \int_0^t \frac{dN^w(u)}{Y^w(u)} - \int_0^t \frac{\sum_{i=1}^{n} Y_i^w(u) d\Lambda_{Pi}(u)}{Y^w(u)}$$

where $N_i^w(t) = N_i(t)/S_{Pi}(t)$, $N^w(t) = \sum_{i=1}^n N_i^w(t)$, $Y_i^w(t) = Y_i(t)/S_{Pi}(t)$ and $Y^w(t) = \sum_{i=1}^n Y_i^w(t)$.

2.1.5. The specific survival. The classical method of specific survival was also used to estimate net survival. The specific survival considers as events only deaths because of cancer. Patients still alive or deceased from other causes are censored. The specific survival was estimated using the Kaplan–Meier estimator and its standard error using Greenwood formula. As mentioned earlier, unlike the other estimators studied here, this method requires the knowledge of the causes of death.

2.2. Parametric estimators based on the additive excess hazard model

Instead of writing the overall survival as the product of the population survival by the net survival (Equation (1)), the overall mortality hazard, λ_o, can be written as $\lambda_e + \lambda_p$; that is the excess hazard (mortality hazard directly or indirectly related to cancer) plus the population mortality hazard [9].

$$\lambda_o(t, a, \mathbf{x}, \mathbf{z}) = \lambda_e(t, a, \mathbf{x}, \mathbf{z}) + \lambda_P(a + t, \mathbf{z}) \quad (3)$$

Here, t is the time elapsed since diagnosis, a the age at diagnosis, $\mathbf{x}$ the vector of covariates, and $\mathbf{z}$ the vector of the variables available in the life table.

2.2.1. The univariable modelling estimator. The strategy proposed by Remontet *et al.* [14] to estimate the excess hazard of a group of patients uses the following model:

$$\log[\lambda_e(t)] = f_0(t)$$

where f_0 is chosen among six candidate functions (a cubic spline with two knots at 1 and 5 years, a cubic spline with one knot at 1 year, a cubic polynomial, a quadratic polynomial, a linear function and a constant function) using the Akaike information criterion. The net survival of the group is then obtained by $\hat{S}_e(t) = \exp\left(-\hat{\Lambda}_e(t)\right)$, where $\hat{\Lambda}_e(t)$ is the cumulative excess hazard. The confidence interval of the estimated net survival is calculated using the Delta method, assuming the normality of $\log(\hat{\Lambda}_e(t))$.

2.2.2. The multivariable modelling estimator. Here, a multivariable proportional hazard regression model that includes covariate age is fitted. The model may be written as

$$\log(\lambda_e(t, age)) = f_0(t) + \beta \cdot age$$

where f_0 is a cubic spline with one knot at one year and β the proportional effect of age on the excess hazard. In this model, $\hat{S}_e(t, age_i)$, the individual survival at time t according to age_i, can be calculated and the net survival of the group estimated by

$$\hat{S}_e(t) = \frac{1}{n} \sum_{i=1}^{n} \hat{S}_e(t, age_i)$$

The 95% confidence interval of the estimated net survival is calculated using Monte Carlo method: Let β^* be the vector of parameters (including the cubic spline coefficients and parameter β). One thousand realizations of β^* can be run by sampling from a multinormal random variable with mean equal to $\hat{\beta}^*$ and variance equal to the variance–covariance matrix $\text{var}(\hat{\beta}^*)$. Thus, the net survival of the group is calculated using the log of the cumulative hazard. The lower and upper limits of the confidence interval are the 2.5th and the 97.5th percentiles of the thousand log-cumulative hazards.

When the above-presented multivariable model fails to converge, other regression models can be fitted: $\log(\lambda_e(t, age)) = f_0(t) + \beta \cdot age$ with $f_0(t)$ now chosen as a cubic polynomial, a quadratic polynomial, a linear, or a constant function using the Akaike information criterion.

3. Simulation design, theoretical parameters and model performance

3.1. Simulation design

The above estimators of net survival were compared through simulations. Covariate *age*, the most influential life-table characteristic, was supposed to cause informative censoring through two different mechanisms: (i) it influences both the population mortality hazard and the excess mortality hazard (which is the case in almost all types of cancer [21]); (ii) it influences the potential follow-up time, with two situations: on average, the potential follow-up time is lower in young than in old patients and; on average, the potential follow-up time is lower in old than in young patients. Thus, six scenarios were simulated (see Table I).

In each scenario, covariate *sex* followed a uniform law that generated as many men as women and covariate *year of diagnosis* followed a uniform law between 1980 and 1990. As shown in Figure 1, the distribution of *age* was considered differently in scenarios 1 and 2 than in scenarios 3 to 6. In the former scenarios, that distribution corresponded approximately to that observed for colon cancer in French cancer incidence data [22]. In the latter scenarios, the distribution of age corresponded to two types of

Table I. Simulated scenarios.

Scenario and follow-up	Age effect on the excess hazard	Early censoring for young patients	Early censoring for old patients	Theoretical net survival (%)	Percent of cancer deaths*	Percent of deaths from other causes*	Censored*
Scenario 1	No	No	No				
5 years				29.6	65.7	10.6	23.7
10 years				14.9	74.8	13.7	11. 5
Scenario 2	Yes	No	No				
5 years				32.6	62.6	8.5	29.0
10 years				20.4	70.5	10.6	18.9
Scenario 3	No	Yes	No				
5 years				29.6	64. 9	12.1	23.1
10 years				14.9	73.3	15.9	10.8
Scenario 4	Yes	Yes	No				
5 years				26.8	67. 5	9.9	22.6
10 years				14.5	74.6	12.2	13.2
Scenario 5	No	No	Yes				
5 years				29.6	64.9	11.8	23.3
10 years				14.9	73.8	15.2	10.9
Scenario 6	Yes	No	Yes				
5 years				22.1	68.2	9.5	22.2
10 years				14.0	76.2	11.9	11.9

*Averaged over 1000 simulated samples.

Scenarios 1 and 2: patients are diagnosed between 1980 and 1990 and 25% of them are aged between 30 and 64, 35% between 65 and 74 and 40% between 75 and 85.

Scenarios 3 and 4: 50% of patients are aged between 70 and 85, and are diagnosed between 1980 and 1985, and 50% of patients are aged between 50 and 85, and are diagnosed between 1985 and 1990.

Scenarios 5 and 6: 50% of patients are aged between 50 and 85, and are diagnosed between 1980 and 1985, and 50% of patients are aged between 70 and 85, and are diagnosed between 1985 and 1990.

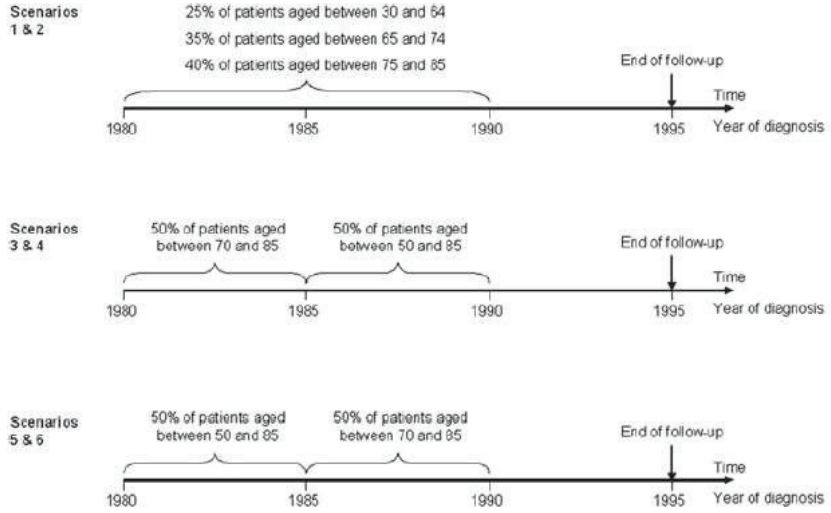


Figure 1. Simulated age distributions at cancer diagnosis in each scenario. In each age class, patients' ages are generated according to a uniform distribution.

situations: (i) scenarios 3 and 4 may illustrate a situation where a standard screening policy would have begun in 1985; (ii) scenarios 5 and 6 correspond to the opposite; it was generated to mimic a long-term cohort study where old patients would be included only after an initial phase of diagnosis, as mentioned by Therneau and Grambsch [23, p 274–275].

The time to each of the two events (T_e for death from cancer and T_P for death from another cause) was generated using the inverse transform method [24]. The log-normal distribution with parameters $\mu = 0.875$ and $\sigma = 1.37$ was used for T_e [25] (which corresponds, for example, to oropharynx cancer with nearly 30% net survival at 5 years). We assumed a proportional effect equal to 0.05 per one year increment and the reference age to be 70 years old (i.e., $\log[\lambda_e(t, age)] = f(t) + 0.05^*(age - 70)$) to simulate the effect of age on T_e. T_P was simulated assuming a yearly piecewise exponential law obtained from the life table of the general population. T_e was assumed to follow a log-normal law because it reflects numerous real-data examples where the cancer mortality hazard increases just after diagnosis for a short time before decreasing continuously thereafter (Figure 2). The common closing date of follow-up was 1995; so, the potential follow-up time C was equal to 1995 minus the year of diagnosis.

The final observation time T was determined as the minimum of $\{T_e, T_P, C\}$, and the data were constructed to: (i) indicate if the subject was censored at $T(\delta = 0)$ or not $(\delta = 1)$ and, in the latter case; (ii) account for the type of event j that occurred at T (this information was used only with the specific survival estimator). In each scenario, we generated $M = 1000$ independent random samples of size $N = 500$.

3.2. *The theoretical net survival*

The theoretical net survival of the group was obtained by averaging the theoretical individual net survivals

$$S_e(t) = \frac{1}{n} \sum_i S_{ei}(t)$$

with $S_{ei}(t) = \exp(-\exp(\beta_{age} \cdot age_i) \cdot \Lambda_{e0}(t))$ and $\Lambda_{e0}(t)$ is the cumulative baseline cancer mortality hazard of the log-normal distribution.

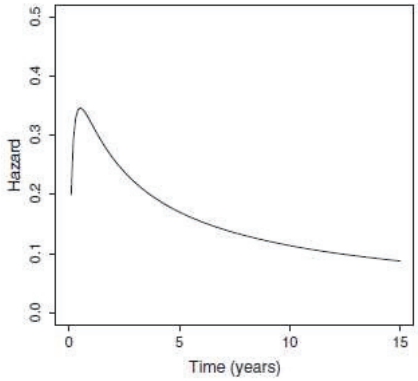


Figure 2. The log-normal hazard with parameters $\mu = 0.875$ and $\sigma = 1.37$.

3.3. Assessment of the performance of the estimators: statistical indicators of performance

The statistical indicators of performance were the bias, the root mean square error (RMSE) and the coverage rate. The bias at time t was the difference between the average of net survival estimated for each sample j, $\hat{S}_{ej}$ and S_e, the theoretical value of the net survival of the whole group: $\frac{1}{M}\sum_{j=1}^{M}(\hat{S}_{ej}(t) - S_e(t))$. The coverage rate was the percentage of confidence intervals that included $S_e(t)$. The RMSE was $\sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(\hat{S}_{ej}(t) - S_e(t)\right)^2}$. These statistical indicators were calculated at $t = 1, 3, 5$ and 10 years.

4. Results

The results presented here were restricted to the net survival estimates at 5 and 10 years of follow-up because at 1 and 3 years, the differences in performance were not sufficiently large to be reported.

With Scenario 1 (no age effect on the excess mortality hazard or on the potential follow-up), all the estimators were unbiased with a coverage rate close to the nominal value of 95% (Table II). The values of the RMSE were close to each other whatever the time at which the net survival was estimated.

With Scenario 2 (presence of an age effect on the excess mortality hazard but not on the potential follow-up), the estimators could be separated into three groups: Group 1 included Ederer I, Hakulinen and the univariable modelling estimators which were all biased (7.04%, 7.12% and 5.40% at 10 years, respectively) and whose coverage rates were far below 95% (19.8%, 19.1% and 36.1% at 10 years, respectively). Group 2 included Ederer II estimator and the specific-survival estimator. These were less biased than Group 1 methods (2.00% and 2.01%, respectively) and their coverage rates were lower than 95% but not as much as the previous methods (88.7% and 86.7%, respectively). Group 3 methods included the multivariable modelling estimator and the Pohar-Perme estimator. These methods showed a good performance, with unbiased estimates (0.06% and 0.19%, respectively) and comparable RMSEs (1.9% and 2.1%, respectively). The coverage rate was good with the multivariable modelling estimator (94.6%), but rather high with the Pohar-Perme estimator (97.8%).

With Scenario 3 (presence of an age effect on the potential follow-up but not on the excess mortality hazard), all estimator biases were close to zero. Only Ederer I estimator departed from the others in terms of bias (−0.73%). The coverage rates were all close to 95%. The RMSEs were all close to 2.4%. Similar trends were observed with Scenario 5.

With scenario 4 (presence of an age effect on the excess mortality hazard and on the potential time of follow-up), the estimators could be classified into two groups. In one group, Ederer II, the specific-survival, the multivariable modelling and Pohar-Perme estimators gave correct estimations of net survival with good coverage rates. In the other group, Ederer I, Hakulinen and the univariable modelling estimators showed important biases and the coverage rates were lower than 95%. With scenario 6, the same classification of the estimators was observed but with more pronounced biases and lower coverage rates for Ederer I, Hakulinen and the univariable modelling estimators than in Scenario 4.

Table II. Performances of the estimators of the net survival at 5 and 10 years over 1000 simulated samples.

	Ederer I	Ederer II	Hakulinen	Univariable model	Multivariable model	Pohar	Specific survival
Scenario 1							
$t = 5$ years							
Bias	−0.18	−0.17	−0.18	0.20	0.26	−0.13	−0.16
RMSE	2.41	2.42	2.41	2.21	2.27	2.49	2.17
Coverage rate	93.5	93.5	93.5	95.2	95.1	94.0	95.2
$t = 10$ years							
Bias	−0.05	−0.13	−0.12	0.01	0.04	0.13	−0.13
RMSE	2.4	2.39	2.39	2.2	2.25	2.82	2.06
Coverage rate	94.8	94.4	94.7	94.9	95.4	94.5	94.4
Scenario 2							
$t = 5$ years							
Bias	3.31	1.21	3.31	4.18	0.23	−0.05	1.24
RMSE	3.97	2.42	3.97	4.71	1.96	2.11	2.39
Coverage rate	74.5	93.9	74.5	57.3	94.5	97.1	92.9
$t = 10$ years							
Bias	7.04	2.0	7.12	5.4	0.06	0.19	2.01
RMSE	7.44	2.83	7.51	5.84	1.87	2.12	2.77
Coverage rate	19.8	88.7	19.1	36.1	94.6	97.8	86.7
Scenario 3							
$t = 5$ years							
Bias	−0.02	−0.02	−0.02	0.35	0.43	0.02	0.01
RMSE	2.33	2.35	2.33	2.27	2.33	2.41	2.14
Coverage rate	95.6	95.1	95.6	94.9	94.9	95.4	95.0
$t = 10$ years							
Bias	−0.73	0.01	0.02	0.36	0.34	0.35	0.0
RMSE	2.37	2.39	2.37	2.29	2.33	2.88	2.03
Coverage rate	96.6	96.3	96.0	95.7	97.0	95.5	95.8
Scenario 4							
$t = 5$ years							
Bias	2.51	0.96	2.51	2.94	0.29	0.04	0.96
RMSE	3.37	2.36	3.37	3.66	2	2.16	2.21
Coverage rate	82.9	94.3	82.9	75.3	95.9	96.8	94.4
$t = 10$ years							
Bias	3.21	0.71	4.1	2.78	0.21	−0.42	0.72
RMSE	4.23	2.47	5.02	3.75	2.11	2.56	2.19
Coverage rate	74.7	94.0	64.4	78.4	95.5	96.5	93.8
Scenario 5							
$t = 5$ years							
Bias	0.08	0.09	0.08	0.52	0.58	0.13	0.09
RMSE	2.41	2.43	2.41	2.31	2.36	2.5	2.21
Coverage rate	94.0	93.9	94.0	94.6	95.2	94.7	95.1
$t = 10$ years							
Bias	0.53	0.03	0.04	0.22	0.26	0.23	0.05
RMSE	2.48	2.35	2.35	2.2	2.29	2.81	2.03
Coverage rate	94.4	95.0	94.9	95.0	94.5	95.2	94.8
Scenario 6							
$t = 5$							
Bias	2.48	1.06	2.48	2.96	0.53	0.21	1.00
RMSE	3.26	2.30	3.26	3.60	1.97	2.05	2.22
Coverage rate	84.7	94.3	84.7	76.8	96.3	97.5	93.6
$t = 10$							
Bias	5.58	1.90	4.94	3.67	0.22	0.82	1.87
RMSE	6.02	2.66	5.40	4.19	1.76	2.04	2.56
Coverage rate	30.9	84.6	38.1	58.1	96.4	96.8	83.6

Note: Survival and the indicators of performance are expressed in percentages.
Scenario 1: no age effect on the excess mortality hazard or on the potential follow-up.
Scenario 2: age effect on the excess mortality hazard but not on the potential follow-up.
Scenario 3: age effect on the potential follow-up (young patients have, on average, a lower potential follow-up times) but not on the excess mortality hazard.
Scenario 4: age effect on the excess mortality hazard and on the potential time of follow-up (young patients have, on average, a lower potential follow-up times).
Scenario 5: age effect on the potential follow-up (old patients have, on average, a lower potential follow-up times) but not on the excess mortality hazard.
Scenario 6: age effect on the excess mortality hazard and on the potential time of follow-up (old patients have, on average, a lower potential follow-up times).

5. Discussion

In the present work, we compared several estimators of net survival. The theoretical reasons for which some of these estimators provide biased estimates of net survival have been recently discussed by Pohar-Perme *et al.* [5]. Our simulation study aimed at illustrating these reasons, quantifying the bias associated with each method, and studying other performance indicators: the RMSE and the coverage rate. Furthermore, the performances of some estimators, especially those based on the multivariable regression approach, were easily and conveniently obtained through this simulation study.

In our scenarios, covariate *age* was the shared factor that could induce informative censoring because it affects the expected mortality and, depending on the scenario, the excess mortality hazard too. In addition, some scenarios allowed studying the effect of age on the potential time of follow-up, which induces informative censoring by heterogeneous withdrawal of patients from observation. In real-life situations, other factors such as sex or year of diagnosis are also known to affect both the excess hazard and the expected hazard inducing thus an informative censoring and, possibly, biased estimates when this censoring is not taken into account.

The performances of the studied estimators at 5 and 10 years are summarized in Table III; they are classified according to three categories 'good', 'unsatisfactory' and 'bad'. As expected, only two methods, the Pohar-Perme estimator and the multivariable modelling estimator, give unbiased estimates of net survival.

Moreover, four points catch our attention. First, as expected by theory, Ederer II and specific survival were very close, showing that, when the other-causes mortality hazard (used for specific survival) is correctly reflected by the population mortality hazard, these two estimators are closely related and estimate the same quantity [5]. Interestingly, even when the 'true' causes of death are known, the estimator of the net survival based on the specific survival would be also biased in the presence of an informative

Table III. Summary of the performances of net survival estimators at 5 and 10 years.

	Ederer I	Ederer II	Hakulinen	Univariable model	Multivariable model	Pohar	Specific survival
Scenario 1							
t = 5 years	Good	Good	Good	Good	Good	Good	Good
t = 10 years	Good	Good	Good	Good	Good	Good	Good
Scenario 2							
t = 5 years	Bad	Unsatisfactory	Bad	Bad	Good	Good	Unsatisfactory
t = 10 years	Bad	Unsatisfactory	Bad	Bad	Good	Good	Unsatisfactory
Scenario 3							
t = 5 years	Good	Good	Good	Good	Good	Good	Good
t = 10 years	Good	Good	Good	Good	Good	Good	Good
Scenario 4							
t = 5 years	Unsatisfactory	Good	Unsatisfactory	Unsatisfactory	Good	Good	Good
t = 10 years	Bad	Good	Bad	Unsatisfactory	Good	Good	Good
Scenario 5							
t = 5 years	Good	Good	Good	Good	Good	Good	Good
t = 10 years	Good	Good	Good	Good	Good	Good	Good
Scenario 6							
t = 5 years	Unsatisfactory	Unsatisfactory	Unsatisfactory	Unsatisfactory	Good	Good	Unsatisfactory
t = 10 years	Bad	Unsatisfactory	Bad	Bad	Good	Good	Unsatisfactory

Note: The category Good is defined when the bias is less than 0.01, Unsatisfactory when the bias is between 0.01 and 0.03 and Bad when the bias is greater than 0.03.

Scenario 1: no age effect on the excess mortality hazard or on the potential follow-up.

Scenario 2: age effect on the excess mortality hazard but not on the potential follow-up.

Scenario 3: age effect on the potential follow-up (young patients have, on average, a lower potential follow-up times) but not on the excess mortality hazard.

Scenario 4: age effect on the excess mortality hazard and on the potential time of follow-up (young patients have, on average, a lower potential follow-up times).

Scenario 5: age effect on the potential follow-up (old patients have, on average, a lower potential follow-up times) but not on the excess mortality hazard.

Scenario 6: age effect on the excess mortality hazard and on the potential time of follow-up (old patients have, on average, a lower potential follow-up times).

censoring [26]. Second, Hakulinen estimator was initially proposed to deal with heterogeneous potential times of follow-up. This estimator had good performance only when there was no age effect on cancer mortality hazard (scenarios 3 or 5), but does not differ from other methods. This suggests that the real impact of informative censoring because of heterogeneous time of follow-up is small (even though the design of the simulation leads to an important change over time of the distribution of age). Third, a smaller bias is observed with Scenario 4 compared with Scenario 6 for all estimators. As previously shown [5], the ratio estimates overestimate the net survival because it approaches the net survival of those who have the best population survival. However, when there were less young people still at risk at end of follow-up (Scenario 4), the Ederer I and Hakulinen estimators underestimated the net survival. Thus, the small bias observed with Scenario 4 for these estimators was the result of opposite effects of two different biases. However, with Scenario 6 where the biases worked in the same direction, the Ederer I and Hakulinen estimators gave more biased results. Fourth, we noticed that the coverage rate of the Pohar-Perme estimator was slightly higher than the nominal value of 95% for scenarios in which an age-effect was simulated; these results need further investigations.

To better test the performance of the Pohar-Perme and the multivariable modelling estimators, we studied the same six scenarios but with a low cancer mortality hazard. The results not shown here led to similar conclusions (see Table III). However, we observed that the bias of the multivariable modelling estimator was slightly higher than that of the Pohar-Perme estimator. One explanation could be that, contrary to the scenarios with high excess mortality hazards, the model with a cubic regression spline with one knot at one year had more difficulty to converge. For example, with Scenario 2, this model failed to converge two times over 1000 simulated samples in the case of high cancer mortality but 189 times in the case of low cancer mortality. Using only the datasets with which the cubic spline converged, the bias of the multivariable modelling estimator was equal to 0.8% instead of 1.1%, but was still higher than the bias of the Pohar-Perme estimator (equal to 0.2%). To make sure that this bias is independent of the sample size, additional analyses with 1000 patients in each dataset were carried out; they led to similar results. Another explanation for the bias of the multivariable modelling estimator could be the positivity constraint imposed on the excess hazard to obtain a genuine survival curve. We did not remove this constraint because the aim of our work was to compare previously published methods [14].

Furthermore, all the methods (except specific survival) are based on the assumption that the expected mortality hazard of a cancer patient is correctly reflected by that of an individual from the general population having the same life-table variables (age, sex...). To evaluate their robustness regarding this assumption, we analyzed two situations where the life-table is inappropriate for a subgroup of patients (i) because of an unmeasured covariate and (ii) because of an unmeasured covariate that influences the excess mortality hazard. These two situations led to bias in all estimators, but the results showed that even in those situations, the Pohar-Perme estimator and the multivariable modelling estimator remained the best estimators of the net survival (results not shown). It must be noted that, although the Pohar-Perme estimator is based on weighting by the inverse of the survival probabilities of appropriate population strata, this estimator does not seem to be more sensitive than the others to the expected hazards obtained using an inappropriate life-table.

The multivariable regression model used in this work reflected clearly the way the data were simulated, that is, with a flexible function for the baseline excess hazard and a proportional and linear effect of age. However, in real-life applications, the analyst is ignorant of the true hazard and the form of the age effect. Furthermore, the multivariable model must include all covariates with potential informative censoring effect (i.e., in practice, variables defining the population life-table). The analyst has then to deal with the complex problem of determining the 'best model' from a set of variables taking into account the time-dependent and nonlinear effects of each one and, possibly, their interactions. Assessing the goodness of fit of a multivariable regression model in the excess hazard setting had been already proposed [27, 28], but the model-building strategy is known to be difficult and is still the object of many research works [14, 29–32]. One asset of the non-parametric Pohar-Perme estimator is that no model-building strategy is needed. However, the multivariable modelling estimator has, of course, some advantages over the nonparametric estimator: it provides smooth estimates of the excess mortality hazard and of covariate effects, its efficiency is usually greater (i.e., the variance of the estimates is smaller), it allows adjusting simultaneously for several covariates, and allows forecasting survival over time.

In conclusion, among the estimators studied here, two may be retained to estimate net survival: Pohar-Perme estimator and the multivariable modelling estimator adjusted for life-table covariates. To take into the account informative censoring, the former uses the inverse probability weighting procedure whereas

the latter adjusts the model on the life-table variables, which are responsible for the informative censoring. It should be mentioned here that the use of these two unbiased estimators of net survival does not exempt from age-standardisation. Indeed, in almost all cancers, net survival depends on age. Thus, comparing net survival between two countries (with two different age distributions) compels the analysts to use age-standardised net survival.

In practice, we advise the use of Pohar-Perme estimator because of its simplicity and its easy implementation (contrary to the multivariable model, which still raises some questions). Moreover, it is available in the *relsurv* R package (function *rs.surv*()). In our opinion, the Pohar-Perme estimator should be preferred in survival studies by cancer registries when only one summary measure is needed for all patients and for each age-group. In addition, these summary measures are consistent with further analyses that may use the multivariable modelling estimator.

Acknowledgements

The authors thank the ANR (Agence Nationale de la Recherche) for supporting the study. The work of the first author was funded by grants from MESURE group (ANR grant number ANR-09-BLAN-0357). The authors are also grateful to M. Pohar-Perme, J. Stare, and J. Estève for their helpful scientific comments and to J. Iwaz for editing the latest drafts of the manuscript.

References

1. Percy C, Stanek E, Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. *American Journal of Public Health* 1981; **71**(3):242–250.
2. Ashworth TG. Inadequacy of death certification: proposal for change. *Journal of Clinical Pathology* 1991; **44**(4):265–268.
3. Aalen O, Borgan Ø, Gjessing H. *Survival and event history analysis*. Springer-Verlag: New York, 2008.
4. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Wiley: New York, 2002.
5. Perme MP, Stare J, Esteve J. On estimation in relative survival. *Biometrics* 2011. DOI: 10.1111/j.1541-0420.2011.01640.x.
6. Buckley JD. Additive and multiplicative models for relative survival rates. *Biometrics* 1984; **40**(1):51–62.
7. Hakulinen T, Tenkanen L. Regression analysis of relative survival rates. *Applied Statistics* 1987; **36**:309–317.
8. Andersen PK, Vaeth M. Simple parametric and nonparametric models for excess and relative mortality. *Biometrics* 1989; **45**(2):523–535.
9. Esteve J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine* 1990; **9**(5):529–538.
10. Ederer F, Heise H. Instructions to ibm 650 programmers in processing survival computations, methodological note no. 10, end results evaluation section. *Technical report, National Cancer Institute, Bethesda MD* 1959.
11. Bolard P, Quantin C, Abrahamowicz M, Esteve J, Giorgi R, Chadha-Boreham H, Binquet C, Faivre J. Assessing time-by-covariate interactions in relative survival models using restrictive cubic spline functions. *Journal of Cancer Epidemiology and Prevention* 2002; **7**(3):113–122.
12. Giorgi R, Abrahamowicz M, Quantin C, Bolard P, Esteve J, Gouvernet J, Faivre J. A relative survival regression model using B-spline functions to model non-proportional hazards. *Statistics in Medicine* 2003; **22**(17):2767–2784.
13. Lambert PC, Smith LK, Jones DR, Botha JL. Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. *Statistics in Medicine* 2005; **24**(24):3871–3885.
14. Remontet L, Bossard N, Belot A, Esteve J. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in Medicine* 2007; **26**(10):2214–2228.
15. Nelson CP, Lambert PC, Squire IB, Jones DR. Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine* 2007; **26**(30):5486–5498.
16. Perme MP, Henderson R, Stare J. An approach to estimation in relative survival regression. *Biostatistics* 2009; **10**(1):136–146.
17. Sasieni PD. Proportional excess hazards. *Biometrika* 1996; **83**(1):127–141.
18. Ederer F, Axtell LM, Cutler SJ. The relative survival rate: a statistical methodology. *National Cancer Institute Monograph* 1961; **6:101-21**.
19. Hakulinen T. Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics* 1982; **38**:933–942.
20. Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemilology: Methodological Issues*, Jewell NP, Dietz K, Farewell VT (eds). Birkhäuser: Boston, 1992; 297–331.
21. Bossard N, Velten M, Remontet L, Belot A, Maarouf N, Bouvier AM, Guizard AV, Tretarre B, Launoy G, Colonna M, Danzon A, Molinie F, Troussard X, Bourdon-Raverdy N, Carli PM, Jaffre A, Bessaguet C, Sauleau E, Schvartz C, Arveux P, Maynadie M, Grosclaude P, Esteve J, Faivre J. Survival of cancer patients in France: a population-based study from The Association of the French Cancer Registries (FRANCIM). *European Journal of Cancer* 2007; **43**(1):149–160.
22. Le Teuff G, Abrahamowicz M, Bolard P, Quantin C. Comparison of Cox's and relative survival models when estimating the effects of prognostic factors on disease-specific mortality: a simulation study under proportional excess hazards. *Statistics in Medicine* 2005; **24**(24):3887–3909.
23. Therneau TM, Grambsch PM. *Modeling Survival Data: extending the Cox Model*. Springer-Verlag: New York, 2000.
24. Ross SM. *Simulation, Fourth Edition*. Elsevier Academic Press: Amsterdam, 2006.
25. Collet D. *Modelling Survival Data in Medical Research. Section 6.1*, 2nd ed. Chapman & Hall/CRC: London, 2003.

26. Robins JM. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. *Proceedings of the American Statistical Association - Biopharmaceutical*, 1993.
27. Cortese G, Scheike TH. Dynamic regression hazards models for relative survival. *Statistics in Medicine* 2008; **27**(18):3563–3584.
28. Stare J, Pohar M, Henderson R. Goodness of fit of relative survival models. *Statistics in Medicine* 2005; **24**(24):3911–3925.
29. Kooperberg C, Stone CJ, Truong YK. Hazard Regression. *Journal of the American Statistical Association* 1995; **90**(429):78–94.
30. Abrahamowicz M, MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in Medicine* 2007; **26**(2):392–408.
31. Sauerbrei W, Royston P, Look M. A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biometrical Journal* 2007; **49**(3):453–473.
32. Sauerbrei W, Royston P. *Multivariable Model Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables*. Wiley: Chichester, 2008.

II.2.6. Evaluation de l'ampleur des erreurs sur l'estimation de la survie nette sur données réelles

II.2.6.a. Illustration des résultats issus des différentes méthodes sur des données réelles

Pour faire suite à la partie précédente dont l'objectif était de comparer les méthodes d'estimation de la survie nette sur données simulées, l'objectif de cette partie est d'illustrer les résultats provenant de ces différentes méthodes sur des données réelles issues de la base commune du réseau des registres français du cancer (base FRANCIM). Plus précisément, il s'agit de quantifier, sur données réelles, l'ampleur des erreurs fournit par les méthodes dites « biaisées » par rapport à la méthode de Pohar-Perme prise comme référence, ainsi que d'étudier l'influence du temps de suivi, du pronostic du cancer, et de la covariable âge sur ces erreurs. En effet, la méthode de Pohar-Perme, récemment développée, a montré qu'elle présentait des avantages majeurs par rapport aux méthodes utilisées jusqu'à présent car elle corrige les effets de la censure informative liée à la mortalité « autres causes ». Il est donc essentiel que la méthodologie proposée pour l'estimation de la survie nette soit reconnue scientifiquement à l'aide d'exploration sur données simulées ainsi que d'explorations sur données réelles. Ces travaux jouent un rôle déterminant pour l'acceptation et l'utilisation de cette méthode : il est important que les registres de cancer abandonnent les méthodes classiques et utilisent la nouvelle méthode de Pohar pour des estimations ponctuelles de la survie nette.

Ce travail a fait l'objet d'un article paru dans *International Journal of Cancer*.

IJC
International Journal of Cancer

Cancer net survival on registry data: Use of the new unbiased Pohar-Perme estimator and magnitude of the bias with the classical methods

Laurent Roche[1,2,3,4], Coraline Danieli[1,2,3,4], Aurélien Belot[1,2,3,4,5], Pascale Grosclaude[6], Anne-Marie Bouvier[7], Michel Velten[8], Jean Iwaz[1,2,3,4], Laurent Remontet[1,2,3,4] and Nadine Bossard[1,2,3,4]

[1] Hospices Civils de Lyon, Service de Biostatistique, F-69003, Lyon, France
[2] Université de Lyon, F-69000, Lyon, France
[3] Université Lyon 1, F-69100, Villeurbanne, France
[4] CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biotatistique-Santé, F-69100, Villeurbanne, France
[5] Institut de Veille Sanitaire, Département des maladies chroniques et traumatismes, F-94410, Saint-Maurice, France
[6] Registre des cancers du Tarn, Institut Claudius Regaud, F-31000, Toulouse, France
[7] Registre bourguignon des cancers digestifs, INSERM U866, Université de Bourgogne, Centre Hospitalier Universitaire de Dijon, F-21000, Dijon, France
[8] Registre des cancers du Bas-Rhin, Laboratoire d'épidémiologie et de santé publique, Université de Strasbourg, F-67085 Strasbourg, France

Net survival, the survival which might occur if cancer was the only cause of death, is a major epidemiological indicator required for international or temporal comparisons. Recent findings have shown that all classical methods used for routine estimation of net survival from cancer-registry data, sometimes called "relative-survival methods," provide biased estimates. Meanwhile, an *unbiased* estimator, the Pohar-Perme estimator (PPE), was recently proposed. Using real data, we investigated the magnitude of the errors made by four "relative-survival" methods (Ederer I, Hakulinen, Ederer II and a univariable regression model) *vs.* PPE as reference and examined the influence of time of follow-up, cancer prognosis, and age on the errors made. The data concerned seven cancer sites (2,51,316 cases) collected by FRANCIM cancer registries. Net survivals were estimated at 5, 10 and 15 years postdiagnosis. At 5 years, the errors were generally small. At 10 years, in good-prognosis cancers, the errors made in nonstandardized estimates with all classical methods were generally great (+2.7 to +9% points in prostate cancer) and increased in age-class estimations (*vs.* 5-year ones). At 15 years, in bad- or average-prognosis cancers, the errors were often substantial whatever the nature of the estimation. In good-prognosis cancers, the errors in nonstandardized estimates of all classical methods were great and sometimes very important. With all classical methods, great errors occurred in age-class estimates resulting in errors in age-standardized estimates (+0.4 to +3.2% points in breast cancer). In estimating net survival, cancer registries should abandon all classical methods and adopt the new Pohar-Perme estimator.

Introduction

Net survival is defined as the survival which might occur if all risks of dying from other causes than the disease of interest, here cancer, were removed.[1] The main idea of net survival is to study the proportion of patients dying from direct or indirect consequences of cancer. Net survival is now a major epidemiological indicator,[1] and already routinely estimated in many countries from data collected by cancer registries.

Two approaches may be adopted to estimate net survival. The first one is the cause-specific approach, which requires knowing the causes of death. The second one uses the all-cause mortality of the study group and the "expected" mortality of a disease-free group having the same demographic characteristics as the study group. Here, the expected mortality is assumed to reflect correctly the mortality due to other causes than cancer and is usually obtained from the general population life tables. The mortality due to cancer is then deduced from the all-cause and other-cause mortalities. The second approach is preferred in epidemiological studies because the causes of death are often unavailable or unreliable.[2,3] Methods for net survival estimation that do not use the information on the cause of death are often referred to by cancer epidemiologists as "relative-survival methods."

Key words: cancer, registries, net survival, relative survival, bias, survival analyses
Abbreviations: APC: average-prognosis cancer; BPC: bad-prognosis cancer; GPC: good-prognosis cancer; ICM: informative censoring mechanism; PPE: Pohar-Perme estimator; UVM: univariable model
Grant sponsors: Agence Nationale de la Recherche (ANR); **Grant number:** ANR-09-BLAN-0357; **Grant sponsors:** Institut National du Cancer (INCA), Institut de Veille Sanitaire (INVS)
DOI: 10.1002/ijc.27830
History: Received 12 Apr 2012; Accepted 9 Aug 2012; Online 10 Sep 2012
Correspondence to: Laurent Roche, Service de Biostatistique des Hospices Civils de Lyon, 162 Avenue Lacassagne, F-69424 Lyon Cedex 03, France, Tel.: (33)-4-72-11-57-55, E-mail: laurent.roche01@chu-lyon.fr

What's new?

"Net survival" refers to the risk of dying from a particular cancer, after all other risks are removed. Unfortunately, due to inherent biases, most of the statistical methods used to estimate net survival are quite inaccurate. In this study, the authors used a new method called the "Pohar-Perme estimator, " (PPE) to analyze data from cancer registries, with various combinations of prognosis and age distribution. They conclude that PPE lacks the biases of the other methods and should become the preferred standard for estimating net survival.

Within the latter setting, many methods have been adopted in national or international cancer-survival studies. Among these, two similar excess-rate regression models[4,5] were used in four studies (England and Wales,[6] France,[7] Spain[8] and ICBP group[9]). In these studies, the cancer mortality rate was modeled by a step or a smooth function that depends only on the time since diagnosis. Ederer I, Ederer II and Hakulinen methods[10,11] are the three other methods widely used in national or international studies (US SEER program,[12,13] EUROCARE,[14] CONCORD,[15] Norway,[16] Finland[17] and NORDCAN[18]).

All these classical methods may produce substantially different estimates of net survival and some may even lead to inconsistencies.[19] Until recently, there was no clear consensus on which method to choose for point estimations of net survival from cancer registry data. Pohar-Perme *et al.*[20] have recently investigated this issue and have *theoretically* shown that, generally, the previously cited methods do not correctly estimate net survival. This can be more easily understood in a competing risk setting. If the risk of dying from cancer and the risk of dying from other causes are dependent, an "informative censoring mechanism" (ICM) occurs; the patients with relatively high risks of dying from other causes will be removed early from the at-risk group whereas they may also have unequal risks of dying from cancer. If this ICM is not taken into account, the estimates relative to the whole population will be close to the survival of the group with the lowest risk of dying from other causes; therefore biased. Any variable that exerts simultaneous effects on cancer mortality and other-cause mortality induces an ICM. Practically, *all* the demographic variables that define life tables (hereafter "life-table variables") may induce an ICM. Pohar-Perme *et al.*[20] further clarified what is actually estimated by Ederer I, Ederer II and Hakulinen methods.

Pohar-Perme *et al.*[20] proposed a new nonparametric estimator [called hereafter the Pohar-Perme estimator (PPE)] and showed that it is an unbiased estimator of net survival, even in the presence of the ICM induced by the life-table variables. The excess-rate regression models can also provide unbiased estimates but only if the cancer mortality rate is modeled as a function that depends on all life-table variables and if these functional dependencies are correctly specified. The model building strategy is known to be difficult and, to our knowledge, there was no satisfactory solution for routine estimation of net survival from cancer registry data for a large variety of tumor sites. Thus, up-to-now, PPE appears to be the only unbiased estimator of net survival available in this context. Furthermore, a recent simulation study[21] pointed out the substantial biases associated with Ederer I and Hakulinen methods, with an excess-rate regression model with only the effect of time since diagnosis modeled, and, to a lesser extent, with Ederer II method.

Our article extends the works of Pohar-Perme *et al.*[20] and Danieli *et al.*[21] It computed, on real data, the magnitude of the errors made with the classical estimators used in cancer registry studies (*i.e.*, Ederer I, Ederer II, Hakulinen and a method derived from the strategy of Remontet *et al.*[5]) when compared to PPE. Three factors that influence the magnitude of these errors are investigated: (*i*) the time elapsed since cancer diagnosis (the longer the follow-up, the longer the ICM will act); (*ii*) the prognosis of cancer (in lethal cancers, most deaths are due to cancer, which reduces the impact of the ICM, whereas, in low-lethal cancers, more deaths may be due to other causes); (*iii*) the effect of the life-table variables (the stronger the effects of these variables on both cancer and other-cause mortalities, the stronger the association between the two mortalities and the stronger the impact of the ICM on net survival estimation). The influence of the third factor will be restricted here to the effect of age. These comparisons are carried out on data from cancer registries (*FRANCIM* network). Seven cancer sites were considered to explore various combinations of prognosis, age distribution and effect of age on net survival.

Material and Methods

Patients

Our study included all patients aged over 15 years registered between January 1, 1989 and December 31, 2004 by 16 participant French cancer registries of *FRANCIM* network. Seven cancer sites were chosen. First, prostate, breast, lung, colon–rectum and head–neck cancers, because they are the five most incident cancers and have various prognoses. Then, Hodgkin disease and thyroid cancer because the former is very frequent in young people and, in both, the effect of age at diagnosis on net survival is important. The administrative censoring date was the January 1, 2008. Data completeness and quality as well as the search for the vital status are described elsewhere.[22,23] Table 1 summarizes the main characteristics of the included patients.

Expected mortality

Unsmoothed expected rates of mortality were first calculated from life tables defined by sex, age, year and Département of residence. These life tables were provided by the Institut

Table 1. Main characteristics of the patients registered by the 16 participant French cancer registries of the network FRANCIM, according to each of the following cancer sites: lung, head and neck, colon–rectum, prostate, breast, thyroid gland and Hodgkin disease

	Lung	Head and neck	Colon–rectum	Prostate	Breast	Thyroid gland	Hodgkin disease
ICD-O3 code[1]	C33, C34	C01 to C06, C09 to C14	C18 to C21	C61	C50	C739	All[2]
Number of cases: *N*	39,063	18,595	64,171	54,087	66,941	6,199	2,260
Sex: *N* (%)							
Male	32,895 (84.2)	16,200 (87.1)	35,082 (54.7)	54,087 (100)	NA	1,332 (21.5)	1,265 (56.0)
Female	6,168 (15.8)	2,395 (12.9)	29,089 (45.3)	NA	66,941 (100)	4,867 (78.5)	995 (44.0)
Median age at diagnosis	66.0	59.0	71.7	71.6	60.6	50.4	38.4
Age at diagnosis quartiles	56.9, 73.5	50.9, 67.5	63.2, 79.4	65.8, 77.5	49.9, 71.2	39.3, 61.6	26.3, 56.6
Age class: *N* (%)							
15–44	1,701 (04.4)	1,576 (08.5)	1,907 (03.0)	–	8,624 (12.9)	2,238 (36.1)	1,372 (60.7)
45–54	6,000 (15.4)	5,126 (27.6)	5,045 (07.9)	–	15,441 (23.1)	1,521 (24.5)	275 (12.2)
15–54	–	–	–	1,326 (02.5)	–	–	–
55–64	9,986 (25.6)	5,723 (30.8)	11,161 (17.4)	9,783 (18.1)	15,607 (23.3)	1,173 (18.9)	183 (08.1)
65–74	12,795 (32.8)	4,126 (22.2)	20,025 (31.2)	22,938 (42.4)	14,909 (22.3)	798 (12.9)	201 (08.9)
75 and older	8,581 (22.0)	2,044 (11.0)	26,033 (40.6)	–	12,360 (18.5)	469 (07.6)	229 (10.1)
75–84	–	–	–	16,181 (29.9)	–	–	–
85 and older	–	–	–	3,859 (07.1)	–	–	–
Period of diagnosis: *N* (%)							
1989–1991	4,747 (12.2)	4,272 (14.4)	8,374 (13.0)	4,982 (07.9)	7,511 (11.2)	501 (08.1)	290 (12.8)
1992–1994	5,165 (13.2)	2,831 (15.2)	10,286 (16.0)	5,297 (09.8)	9,643 (14.4)	591 (09.5)	293 (13.0)
1995–1997	7,111 (18.2)	3,505 (18.8)	12,758 (19.9)	7,704 (14.2)	12,633 (18.9)	975 (15.7)	436 (19.3)
1998–2000	8,744 (22.4)	4,166 (22.4)	13,643 (21.3)	11,970 (22.1)	14,398 (21.5)	1,541 (24.9)	477 (21.1)
2001–2004	13,296 (34.0)	5,421 (29.2)	19,110 (29.8)	24,834 (45.9)	22,756 (34.0)	2,591 (41.8)	764 (33.8)
Median follow-up (years)	0.8	2.0	3.4	4.7	6.5	6.8	6.9

[1]Hematological codes are always excluded from solid tumor sites. [2]Morphology code for Hodgkin disease: (≥9,650/3 and ≤9,655/3) or (≥9,661/3 and ≤9,667/3).

National de la Statistique et des Etudes Economiques. The expected mortality rates of patients over 30 were smoothed by sex and Département using a Poisson regression model that included a bidimensional spline of year and age (*mgcv* package in R software).[24]

Net survival estimations at 5, 10 and 15 years postdiagnosis

Net survival was estimated at 5, 10 and 15 years postdiagnosis using five methods: Ederer I, Ederer II, Hakulinen, Pohar-Perme and a univariable model (UVM). Age-standardized net survival estimates were obtained with the International standard for cancer-survival weights.[25] Estimations carried out with Ederer I,[10] Ederer II[10] and Hakulinen methods[11] were performed as usual and used individual data.

The fourth method, PPE,[20] may be viewed as a weighted version of Ederer II method. It relies on the inverse probability weighting procedure[26] that allows correcting the bias due to the ICM induced by the life-table variables.

The fifth method, the strategy proposed by Remontet *et al.*,[5] is an excess-rate regression model and was adapted here to estimate net survival at 5, 10 and 15 years postdiagnosis. In this strategy, the observed mortality rate λ_O may be written as the sum of the expected mortality rate plus cancer mortality rate; $\lambda_O = \lambda_E + \lambda_C$.[4] The logarithm of λ_C was modeled as a smoothed parametric function of time chosen according to the Akaike Information Criterion among two cubic regression splines (with two knots at 1 and 5 year, or a single knot at 1 year) and four polynomials (constant, linear, quadratic and cubic). In addition to these six candidate functions, a three-knot cubic regression spline (with knots at 1, 5 and 10 years) was also considered for the analyses at 15 years. The net survival at a given time t was then estimated by $\exp(-\int_0^t \hat{\lambda}_C(s)ds)$. This will be referred to as the UVM.

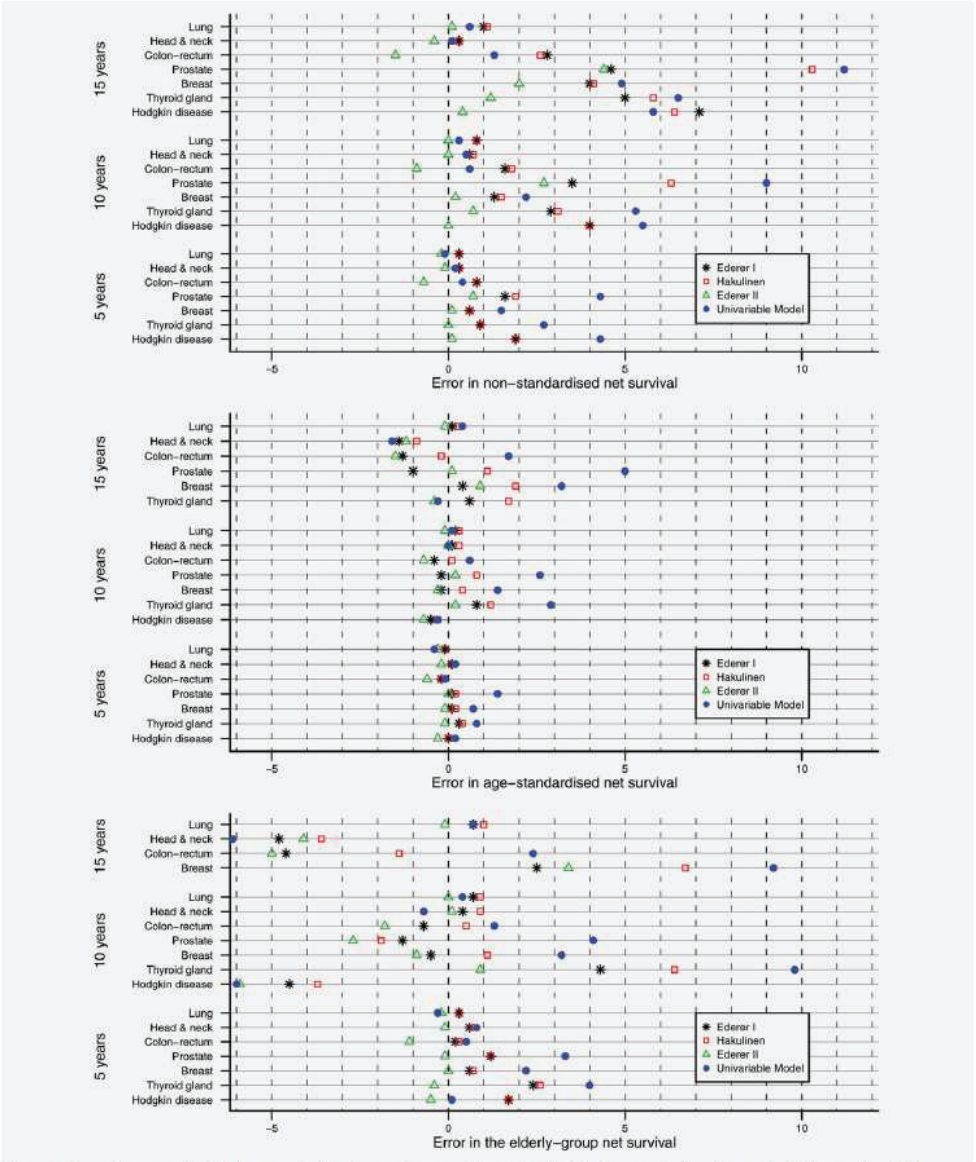


Figure 1. Errors in nonstandardized net survival estimates (top panel), age-standardized net survival estimates (middle panel), and the elderly group net survival estimates (bottom panel) with Ederer I, Hakulinen, Ederer II and the Univariable at 5, 10 and 15 years after diagnosis. A unit on the *x*-axis is a 1% point difference of an estimate obtained with a classical method minus the Pohar-Perme estimate. The elderly group is defined as the group of patients aged over 85 for prostate cancer and over 75 for the six other sites. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 2. Estimates of 5-year net survival (age-standardized or not) obtained with Ederer I, Hakulinen, Ederer II, the UVM and the PPE

	5-Year net survival				
Cancer site	Pohar-Perme	Ederer I	Hakulinen	Ederer II	Univariable model
Lung					
15–44	21.0 [19.1;23.1]	20.9 [19.0;22.9]	20.9 [19.0;22.9]	20.9 [19.0;22.9]	20.3 [18.5;22.3]
45–54	17.2 [16.2;18.2]	17.0 [16.0;18.0]	17.0 [16.0;18.0]	17.0 [16.0;18.0]	16.8 [15.8;17.8]
55–64	16.3 [15.6;17.1]	16.1 [15.4;16.9]	16.1 [15.4;16.9]	16.1 [15.3;16.9]	15.8 [15.1;16.5]
65–74	14.0 [13.3;14.7]	13.7 [13.1;14.4]	13.7 [13.1;14.4]	13.7 [13.0;14.3]	13.7 [13.1;14.3]
75 and older	8.0 [7.2;8.7]	8.3 [7.6;9.1]	8.3 [7.6;9.1]	7.8 [7.2;8.6]	7.7 [7.1;8.4]
Nonstandardized	14.1 [13.7;14.5]	14.4 [14.0;14.8]	14.4 [14.0;14.8]	13.9 [13.5;14.3]	14.0 [13.6;14.3]
Age-standardized	13.7 [13.3;14.0]	13.6 [13.2;14.0]	13.6 [13.2;14.0]	13.4 [13.1;13.8]	13.3 [12.9;13.6]
Head and neck					
15–44	42.2 [39.7;44.7]	42.0 [39.5;44.5]	42.0 [39.5;44.5]	42.0 [39.5;44.5]	41.6 [39.2;44.0]
45–54	37.4 [36.1;38.9]	37.3 [36.0;38.7]	37.3 [36.0;38.7]	37.3 [35.9;38.7]	37.4 [36.0;38.8]
55–64	33.9 [32.6;35.2]	33.7 [32.4;35.1]	33.8 [32.5;35.1]	33.7 [32.4;35.0]	33.6 [32.3;34.9]
65–74	31.4 [29.9;33.1]	31.3 [29.8;32.9]	31.3 [29.8;33.0]	31.2 [29.7;32.9]	31.6 [30.2;33.1]
75 and older	26.4 [23.9;29.2]	27.0 [24.5;29.7]	27.1 [24.6;29.8]	26.3 [23.8;28.9]	27.2 [24.9;29.5]
Nonstandardized	34.2 [33.5;35.0]	34.5 [33.8;35.3]	34.5 [33.8;35.3]	34.1 [33.4;34.8]	34.4 [33.7;35.2]
Age-standardized	32.0 [31.0;33.0]	32.1 [31.1;33.1]	32.1 [31.2;33.1]	31.8 [30.9;32.8]	32.2 [31.3;33.1]
Colon–rectum					
15–44	66.0 [63.8;68.2]	65.9 [63.7;68.1]	65.9 [63.7;68.1]	65.9 [63.6;68.1]	65.5 [63.3;67.6]
45–54	63.7 [62.3;65.1]	63.4 [62.0;64.8]	63.4 [62.0;64.8]	63.4 [62.0;64.8]	63.3 [61.9;64.6]
55–64	62.1 [61.1;63.1]	61.7 [60.8;62.7]	61.8 [60.8;62.7]	61.7 [60.7;62.7]	61.8 [60.8;62.7]
65–74	58.6 [57.8;59.4]	58.2 [57.4;59.0]	58.2 [57.4;59.0]	58.1 [57.3;58.9]	58.5 [57.7;59.2]
75 and older	49.5 [48.6;50.5]	49.7 [48.8;50.6]	49.8 [48.9;50.7]	48.4 [47.6;49.3]	50.0 [49.1;50.8]
Nonstandardized	56.2 [55.7;56.7]	57.0 [56.5;57.5]	57.0 [56.6;57.5]	55.5 [55.1;56.0]	56.6 [56.2;57.1]
Age-standardized	57.9 [57.4;58.4]	57.7 [57.2;58.2]	57.7 [57.3;58.2]	57.3 [56.8;57.8]	57.8 [57.4;58.3]
Prostate					
15–54	83.7 [81.4;86.1]	83.7 [81.2;85.9]	83.7 [81.2;85.9]	83.7 [81.2;86.0]	84.2 [81.9;86.2]
55–64	90.3 [89.4;91.1]	90.2[89.4;91.0]	90.3 [89.4;91.1]	90.3 [89.4;91.1]	90.9 [90.1;91.6]
65–74	88.9 [88.2;89.6]	88.8 [88.1;89.5]	88.9 [88.2;89.6]	88.9 [88.2;89.5]	90.0 [89.5;90.6]
75–84	78.7 [77.5;79.9]	79.3 [78.1;80.4]	79.4 [78.3;80.6]	78.9 [77.7;80.0]	81.6 [80.5;82.7]
85 and older	58.4 [54.7;62.4]	59.6 [56.0;63.4]	59.6 [56.0;63.4]	58.3 [54.8;62.0]	61.7 [58.2;65.1]
Nonstandardized	83.7 [83.2;84.3]	85.3 [84.7;85.8]	85.6 [85.0;86.1]	84.4 [83.8;84.9]	88.0 [87.6;88.4]
Age-standardized	84.1 [83.4;84.7]	84.2 [83.6;84.8]	84.3 [83.7;84.9]	84.1 [83.4;84.7]	85.5 [84.9;86.1]
Breast					
15–44	86.1 [85.3;86.9]	86.1 [85.3;86.8]	86.1 [85.3;86.8]	86.1 [85.3;86.8]	85.9 [85.2;86.6]
45–54	89.8 [89.3;90.3]	89.8 [89.2;90.3]	89.8 [89.2;90.3]	89.8 [89.2;90.3]	89.7 [89.2;90.2]
55–64	88.1 [87.5;88.7]	88.1 [87.5;88.6]	88.1 [87.5;88.6]	88.1 [87.5;88.6]	88.0 [87.5;88.6]
65–74	85.9 [85.2;86.6]	85.8 [85.1;86.5]	85.8 [85.1;86.5]	85.8 [85.1;86.5]	86.2 [85.5;86.8]
75 and older	75.9 [74.5;77.3]	76.5 [75.2;77.8]	76.6 [75.3;77.9]	75.9 [74.6;77.2]	78.1 [77.0;79.2]
Nonstandardized	85.5 [85.1;85.9]	86.1 [85.8;86.5]	86.1 [85.8;86.5]	85.6 [85.3;86.0]	87.0 [86.7;87.3]
Age-standardized	84.0 [83.5;84.5]	84.1 [83.7;84.6]	84.2 [83.7;84.6]	83.9 [83.5;84.4]	84.7 [84.3;85.1]
Thyroid gland					
15–44	99.4 [99.0;99.9]	99.4 [98.8;99.7]	99.4 [98.8;99.7]	99.4 [98.8;99.7]	99.4 [98.8;99.7]
45–54	98.3 [97.4;99.2]	98.3 [97.2;99.1]	98.3 [97.2;99.1]	98.3 [97.2;99.1]	98.6 [97.4;99.2]

Table 2. Estimates of 5-year net survival (age-standardized or not) obtained with Ederer I, Hakulinen, Ederer II, the UVM and the PPE (Continued)

	5-Year net survival				
Cancer site	Pohar-Perme	Ederer I	Hakulinen	Ederer II	Univariable model
55–64	94.8 [93.2;96.5]	94.9 [93.1;96.4]	94.9 [93.1;96.4]	94.8 [93.0;96.3]	95.1 [93.3;96.4]
65–74	84.7 [81.5;87.9]	84.5 [81.1;87.5]	84.5 [81.1;87.6]	84.2 [80.9;87.3]	85.3 [82.1;87.9]
75 and older	57.3 [51.1;64.2]	59.7 [53.4;66.2]	59.9 [53.6;66.5]	56.9 [50.9;63.1]	61.3 [55.5;66.6]
Nonstandardized	93.2 [92.4;94.1]	94.1 [93.3;94.9]	94.1 [93.3;94.9]	93.2 [92.4;93.9]	95.9 [95.3;96.4]
Age-standardized	89.4 [88.2;90.6]	89.7 [88.5;90.9]	89.8 [88.6;90.9]	89.3 [88.1;90.4]	90.2 [89.2;91.2]
Hodgkin disease					
15–44	93.9 [92.6;95.3]	93.8 [92.4;95.1]	93.8 [92.4;95.1]	93.8 [92.4;95.1]	94.1 [92.7;95.2]
45–54	82.9 [78.2;87.9]	82.6 [77.3;87.1]	82.6 [77.3;87.1]	82.6 [77.3;87.0]	82.1 [77.1;86.1]
55–64	78.4 [71.8;85.6]	77.7 [70.3;84.0]	77.7 [70.3;84.0]	77.6 [70.2;83.9]	79.1 [71.8;84.7]
65–74	55.4 [47.9;64.0]	55.1 [47.3;63.1]	55.2 [47.3;63.1]	55.1 [47.2;63.0]	56.4 [48.6;63.4]
75 and older	32.4 [25.1;41.8]	34.1 [26.5;43.0]	34.1 [26.5;43.1]	31.9 [24.8;40.3]	32.5 [25.2;39.9]
Nonstandardized	81.8 [80.0;83.7]	83.7 [81.8;85.4]	83.7 [81.8;85.4]	81.9 [80.0;83.6]	86.1 [84.5;87.5]
Age-standardized	81.3 [79.6;82.9]	81.3 [79.6;82.9]	81.3 [79.6;82.9]	81.0 [79.4;82.7]	81.5 [80.0;83.0]

The seven cancer sites under study were split into three prognostic groups: (*i*) bad-prognosis cancers (BPCs) include lung and head–neck cancers; (*ii*) average-prognosis cancers (APCs) include colon–rectum cancers; and (*iii*) good-prognosis cancers (GPCs) include Hodgkin disease and thyroid gland, prostate and breast cancers.

Results

PPE being the reference, Figure 1 shows the differences between the estimates provided by PPE and those provided by the other methods. Nonstandardized, age-standardized, and the elderly group estimate differences are shown according to the follow-up time and the cancer site. These sites are grossly presented in increasing order of severity. The elderly group is defined as the group of patients aged over 85 for prostate cancer and over 75 for the six other sites.

Results at 5 years postdiagnosis

Table 2 shows the age-classes, nonstandardized and age-standardized estimates of the net survival at 5 years postdiagnosis. For BPCs and APCs, the estimates provided by each of the classical methods are close to PPE estimates. The absolute values of the differences are all less than 0.8% point for nonstandardized estimates and less than 0.6% point for age-standardized estimates (Fig. 1).

For GPCs, higher nonstandardized estimations were obtained with Ederer I, Hakulinen and the UVM than with PPE, especially for prostate cancer (83.7 for PPE *vs.* 85.3, 85.6 and 88 for Ederer I, Hakulinen and the UVM, respectively) and Hodgkin disease (81.8 *vs.* 83.7, 83.7 and 86.1, respectively). The age-standardized estimates differed only slightly, except with the UVM in prostate cancer (+1.4% points). Age-class estimates also differed from PPE estimates in GPCs, but only in the elderly groups: up to +2.4% points for Ederer I (thyroid gland), +2.6 for Hakulinen (thyroid gland) and +4.3 for the UVM (prostate). There were no important differences in any estimation with Ederer II.

Results at 10 years postdiagnosis

Table 3 shows the age-classes, nonstandardized and age-standardized estimates of the net survival at 10 years postdiagnosis. For BPCs, the differences between the estimates according to various methods were small. For APCs, nonstandardized estimates (*vs.* PPE estimates) were 1.6, 1.8 and 0.6% points higher with, respectively, Ederer I, Hakulinen and the UVM but 0.9% point lower with Ederer II (Fig. 1). The differences in age-standardized estimates ranged from −0.7 (Ederer II) to +0.6 (UVM). The age-class estimates did not vary significantly.

For GPCs, whatever the cancer site, the differences (Fig. 1) in nonstandardized estimates became marked at 10 years with Ederer I, Hakulinen and the UVM (prostate: +3.5, +6.3 and +9% points, respectively; Hodgkin disease: +4, +4 and +5.5% points, respectively). There were significant differences between PPE and Ederer II estimations regarding prostate cancer (+2.7% points). The age-standardized estimates provided by Ederer I, Hakulinen and Ederer II methods ranged within 1% point difference with regard to PPE, except for Hakulinen method in thyroid gland cancers (+1.2% points). The age-standardized estimates provided by the UVM were more than 2.5% points higher than PPE estimates for prostate and thyroid gland cancers. The differences in survival estimates by age-class between PPE and the other methods were much higher than those found at 5 years, especially in the elderly groups (for example, see thyroid gland cancer and Hodgkin disease).

Table 3. Estimates of 10-year net survival (age-standardized or not) obtained with Ederer I, Hakulinen, Ederer II, the UVM and the PPE

	10-Year net survival				
Cancer site	Pohar-Perme	Ederer I	Hakulinen	Ederer II	Univariable model
Lung					
15–44	16.5 [14.7;18.6]	16.5 [14.6;18.5]	16.5 [14.6;18.5]	16.4 [14.5;18.4]	16.6 [14.7;18.5]
45–54	12.7 [11.8;13.8]	12.6 [11.6;13.6]	12.6 [11.7;13.6]	12.6 [11.6;13.6]	12.5 [11.6;13.5]
55–64	10.3 [09.6;11.1]	10.2 [9.5;11.0]	10.3 [9.5;11.0]	10.2 [09.5;11.0]	10.2 [9.5;10.9]
65–74	8.2 [7.6;9.0]	8.1 [7.4;8.8]	8.2 [7.5;8.9]	8.1 [7.4;8.8]	8.3 [7.6;9.0]
75 and older	4.8 [3.8;6.0]	5.5 [4.6;6.7]	5.7 [4.7;7.0]	4.8 [4.0;5.9]	5.2 [4.5;6.1]
Nonstandardized	9.1 [8.7;9.5]	9.9 [9.5;10.3]	9.9 [9.5;10.3]	9.1 [8.7;9.5]	9.4 [9.1;9.8]
Age-standardized	8.8 [8.4;9.3]	9.0 [8.5;9.4]	9.1 [8.6;9.5]	8.7 [8.3;9.2]	8.9 [8.5;9.3]
Head and neck					
15–44	28.6 [26.1;31.3]	28.5 [26.0;31.1]	28.5 [26.0;31.2]	28.5 [26.0;31.1]	29.1 [26.7;31.6]
45–54	24.6 [23.2;26.0]	24.5 [23.1;25.9]	24.5 [23.1;26.0]	24.5 [23.1;25.9]	24.5 [23.1;25.9]
55–64	20.2 [18.9;21.5]	20.1 [18.8;21.5]	20.3 [19.0;21.6]	20.1 [18.8;21.5]	20.2 [18.9;21.5]
65–74	18.8 [17.1;20.6]	18.9 [17.2;20.7]	19.0 [17.4;20.8]	18.8 [17.1;20.5]	19.4 [17.8;21.0]
75 and older	17.1 [12.7;23.0]	17.5 [14.2;21.6]	18.0 [14.5;22.1]	17.2 [13.9;21.2]	16.4 [13.4;19.7]
Nonstandardized	21.4 [20.6;22.3]	22.0 [21.3;22.9]	22.1 [21.3;22.9]	21.4 [20.7;22.2]	21.9 [21.1;22.6]
Age-standardized	20.0 [18.5;21.7]	20.1 [18.9;21.4]	20.3 [19.1;21.6]	20.0 [18.8;21.2]	20.0 [18.9;21.1]
Colon–rectum					
15–44	59.5 [57.1;62.0]	59.4 [56.9;61.8]	59.4 [57.0;61.9]	59.4 [56.9;61.8]	59.6 [57.1;62.0]
45–54	56.5 [54.9;58.1]	56.2 [54.6;57.9]	56.3 [54.7;57.9]	56.2 [54.6;57.9]	56.6 [54.9;58.1]
55–64	54.4 [53.3;55.6]	54.1 [52.9;55.3]	54.3 [53.2;55.5]	54.2 [53.0;55.3]	54.8 [53.6;55.9]
65–74	51.2 [50.2;52.3]	51.0 [49.9;52.0]	51.2 [50.2;52.3]	51.0 [49.9;52.0]	51.8 [50.8;52.8]
75 and older	44.8 [42.6;47.1]	44.1 [42.6;45.7]	45.3 [43.8;46.9]	43.0 [41.5;44.5]	46.1 [44.7;47.4]
Nonstandardized	49.9 [48.9;50.8]	51.5 [50.9;52.2]	51.7 [51.0;52.4]	49.0 [48.4;49.6]	50.5 [49.9;51.1]
Age-standardized	51.3 [50.5;52.1]	50.9 [50.3;51.6]	51.4 [50.8;52.1]	50.6 [50.0;51.3]	51.9 [51.3;52.6]
Prostate					
15–54	71.7 [67.4;76.2]	71.5 [67.0;75.8]	71.5 [67.0;75.8]	71.7 [67.1;75.9]	71.2 [66.7;75.2]
55–64	82.6 [81.0;84.2]	81.9 [80.3;83.4]	82.7 [81.1;84.2]	82.6 [81.0;84.2]	83.8 [82.4;85.2]
65–74	78.6 [77.2;79.9]	77.9 [76.5;79.2]	79.0 [77.6;80.3]	78.7 [77.4;80.0]	81.1 [80.1;82.0]
75–84	61.3 [58.6;64.2]	62.2 [59.7;64.7]	64.5 [62.0;67.2]	62.4 [59.9;64.9]	67.6 [65.1;70.0]
85 and older	32.1 [24.0;42.9]	30.8 [23.5;40.3]	30.2 [23.0;39.5]	29.4 [22.4;38.5]	36.2 [25.8;46.7]
Nonstandardized	70.1 [68.6;71.6]	73.6 [72.6;74.6]	76.4 [75.3;77.4]	72.8 [71.8;73.8]	79.1 [78.2;79.9]
Age-standardized	71.4 [70.1;72.7]	71.2 [69.9;72.4]	72.2 [70.9;73.4]	71.6 [70.3;72.8]	74.0 [72.8;75.3]
Breast					
15–44	74.8 [73.7;75.9]	74.7 [73.6;75.8]	74.7 [73.6;75.8]	74.7 [73.6;75.8]	74.6 [73.5;75.6]
45–54	82.7 [81.9;83.5]	82.6 [81.9;83.4]	82.7 [81.9;83.4]	82.6 [81.9;83.4]	82.7 [81.9;83.4]
55–64	79.7 [78.9;80.6]	79.6 [78.7;80.4]	79.7 [78.8;80.6]	79.7 [78.8;80.6]	80.1 [79.2;80.9]
65–74	76.3 [75.2;77.4]	76.1 [75.0;77.2]	76.4 [75.3;77.6]	76.3 [75.1;77.4]	77.7 [76.6;78.7]
75 and older	64.8 [61.6;68.1]	64.3 [62.0;66.6]	65.9 [63.6;68.3]	63.9 [61.6;66.2]	68.0 [65.9;70.0]
Nonstandardized	76.3 [75.6;77.0]	77.6 [77.1;78.1]	77.8 [77.3;78.3]	76.5 [76.0;77.0]	78.5 [78.1;78.9]
Age-standardized	74.4 [73.4;75.5]	74.2 [73.4;74.9]	74.8 [74.0;75.6]	74.1 [73.4;74.9]	75.8 [75.1;76.5]
Thyroid gland					
15–44	99.2 [98.4;99.9]	99.1 [98.2;99.7]	99.1 [98.2;99.7]	99.1 [98.2;99.7]	99.3 [98.2;99.7]
45–54	96.2 [94.5;98.0]	96.2 [94.3;97.8]	96.2 [94.3;97.8]	96.2 [94.3;97.8]	96.8 [94.8;98.1]

Table 3. Estimates of 10-year net survival (age-standardized or not) obtained with Ederer I, Hakulinen, Ederer II, the UVM and the PPE (Continued)

	10-Year net survival				
Cancer site	Pohar-Perme	Ederer I	Hakulinen	Ederer II	Univariable model
55–64	91.8 [89.1;94.6]	92.0 [89.1;94.5]	92.2 [89.3;94.8]	91.9 [89.0;94.4]	95.1 [93.3;96.4]
65–74	78.0 [72.9;83.6]	78.7 [73.5;83.8]	79.2 [73.9;84.2]	78.2 [73.0;83.2]	81.7 [76.2;85.9]
75 and older	38.7 [25.0;59.7]	43.0 [32.1;56.5]	45.1 [33.6;59.3]	39.6 [29.6;52.1]	48.5 [35.1;60.5]
Nonstandardized	89.9 [88.3;91.6]	92.8 [91.5;93.9]	93.0 [91.8;94.2]	90.6 [89.4;91.8]	95.2 [94.3;95.9]
Age-standardized	84.4 [81.8;87.1]	85.2 [83.1;87.3]	85.6 [83.5;87.8]	84.6 [82.6;86.6]	87.3 [85.3;89.5]
Hodgkin disease					
15–44	90.3 [88.4;92.2]	90.2 [88.1;92.0]	90.2 [88.2;92.0]	90.2 [88.2;92.0]	90.7 [88.8;92.2]
45–54	76.1 [70.1;82.6]	76.0 [69.3;81.8]	76.0 [69.3;81.8]	75.8 [69.2;81.7]	74.6 [67.8;80.2]
55–64	67.2 [58.3;77.5]	66.6 [57.0;75.7]	67.0 [57.3;76.1]	66.4 [56.8;75.5]	66.9 [56.8;75.1]
65–74	43.8 [34.6;55.5]	44.3 [34.7;55.0]	44.6 [34.9;55.4]	43.8 [34.3;54.5]	46.4 [36.1;56.1]
75 and older	30.3 [18.4;49.6]	25.8 [14.3;44.9]	26.6 [14.7;46.3]	24.3 [13.4;42.4]	24.3 [15.6;34.1]
Nonstandardized	76.7 [74.3;79.1]	80.7 [78.4;83.0]	80.7 [78.4;83.0]	76.7 [74.5;78.9]	82.2 [80.2;84.0]
Age-standardized	75.9 [73.5;78.4]	75.4 [73.0;77.9]	75.5 [73.1;78.1]	75.2 [72.8;77.6]	75.6[73.6;77.7]

Results at 15 years postdiagnosis

Table 4 shows the age-classes, nonstandardized and age-standardized estimates of the net survival at 15 years postdiagnosis. For BPCs, the nonstandardized estimates were generally slightly higher than PPE estimates. The age-standardized estimates did not markedly differ from PPE for lung cancer and were lower than PPE estimates for head–neck cancer. The age-classes estimates differed from PPE estimates mainly in the elderly group: the former were slightly higher for lung cancer, and lower for head–neck cancer (from −6.1 to −3.6% points).

For APCs, the nonstandardized estimates differences were +2.8, +2.6 and +1.3% points with Ederer I, Hakulinen and the UVM, respectively. The age-standardized estimates differed by −1.3, −0.2 and +1.7% points, respectively, and the estimates in the elderly group by −4.6, −1.4 and 2.4% points, respectively. The estimates obtained with Ederer II were generally lower than those obtained with PPE.

For GPCs, the nonstandardized estimates obtained with Ederer I, Hakulinen and the UVM were systematically much higher than those obtained with PPE: these differences ranged from +4 to +11.2% points. The corresponding Ederer II estimates were systematically higher than PPE estimates (from +0.4 to +4.4% points). For Hodgkin disease, prostate cancer and thyroid gland cancer (not breast cancer), the estimations in the elderly groups were unstable (very wide 95% confidence intervals) because very few persons remain at risk at 15 years; thus, no relevant conclusion can be drawn from these estimations. The estimations in other age-classes were sometimes unstable in thyroid gland and Hodgkin disease for the same reason. The instability of these estimates impacts directly the age-standardized estimation, and reduces its reliability. For breast cancer, the age-standardized estimates obtained with Ederer I, Hakulinen, Ederer II and the UVM were slightly higher than those obtained with PPE (+0.4, +1.9, +0.9 and +3.2% points, respectively). The estimates relative to the elderly group were substantially higher than those obtained with PPE (+2.5, +6.7, +3.4 and +9.2% points, respectively). In the three other sites, differences between estimates were also observed in age-classes that did not include the elderlies (*e.g.*, prostate or thyroid gland cancer).

Discussion

In our article, we point out the magnitude of the errors made when net survival is estimated using the classical Ederer I, Ederer II, Hakulinen, and the UVM instead of PPE. These errors may be very important in nonstandardized estimations of net survival. Great errors may occur in estimations by age-class, essentially in the elderly groups, which generates errors in age-standardized estimations. The classical methods overestimate generally the net survival because patients with low (vs. high) other-cause mortalities die usually less frequently from cancer. Our results are in agreement with the findings of Pohar-Perme *et al.*[20] and Danieli *et al.*[21]

The errors were generally small in BPCs and important in GPCs, especially with the UVM. In BPCs, most deaths occur very early after diagnosis; the all-cause mortality is then essentially influenced by cancer mortality whereas, in GPCs, the all-cause mortality is more influenced by mortality from other causes. This increases the censoring due to death from other causes than cancer (*i.e.*, the ICM). The errors are amplified with the time elapsed since diagnosis because the at-risk person set is more and more distorted in a nonrandom way. At 5 years postdiagnosis, the errors of the estimation with Ederer I, Hakulinen, and the UVM in GPCs were of a few % points. At 15 years postdiagnosis, the bias with Ederer I, Hakulinen, the UVM and Ederer II was often large and sometimes extremely

Table 4. Estimates of 15-year net survival (age-standardized or not) obtained with Ederer I, Hakulinen, Ederer II, the UVM and the PPE

	15-Year net survival				
Cancer site	Pohar-Perme	Ederer I	Hakulinen	Ederer II	Univariable model
Lung					
15–44	13.3 [11.3;15.6]	13.3 [11.3;15.6]	13.3 [11.3;15.7]	13.2 [11.2;15.5]	13.4 [11.3;15.7]
45–54	8.4 [7.2;9.9]	8.3 [7.1;9.8]	8.4 [7.1;9.8]	8.3 [7.1;9.7]	8.7 [7.6;9.9]
55–64	7.3 [6.4;8.2]	7.1 [6.3;8.0]	7.3 [6.4;8.2]	7.2 [6.3;8.1]	7.2 [6.4;8.1]
65–74	5.0 [4.0;6.2]	4.7 [3.8;5.8]	4.8 [3.9;5.9]	4.7 [3.8;5.8]	5.4 [4.5;6.3]
75 and older	3.0 [1.6;5.8]	3.7 [2.0;6.6]	4.0 [2.2;7.4]	2.9 [1.6;5.3]	3.7 [0.8;10.2]
Nonstandardized	6.1 [5.5;6.7]	7.1 [6.6;7.7]	7.2 [6.6;7.7]	6.2 [5.7;6.7]	6.7 [6.3;7.2]
Age-standardized	5.9 [5.3;6.7]	6.0 [5.3;6.8]	6.2 [5.4;7.1]	5.8 [5.2;6.5]	6.3 [5.0;7.8]
Head and neck					
15–44	19.9 [17.0;23.2]	19.9 [17.0;23.2]	19.9 [17.0;23.2]	19.8 [16.9;23.1]	19.9 [17.3;22.8]
45–54	16.6 [15.0;18.5]	16.6 [14.9;18.4]	16.6 [15.0;18.5]	16.6 [14.9;18.4]	16.5 [14.9;18.1]
55–64	12.4 [10.9;14.1]	12.2 [10.7;13.8]	12.4 [10.9;14.1]	12.2 [10.8;13.9]	12.3 [10.9;13.7]
65–74	9.8 [07.8;12.3]	9.9 [07.9;12.3]	10.0 [8.0;12.4]	9.7 [7.7;12.0]	10.5 [8.7;12.5]
75 and older	21.3 [10.6;42.5]	16.5 [10.4;26.2]	17.7 [11.1;28.1]	17.2 [10.8;27.3]	15.2 [10.4;20.8]
Nonstandardized	14.0 [12.7;15.3]	14.3 [13.3;15.3]	14.3 [13.4;15.4]	13.6 [12.7;14.6]	14.1 [13.3;15.0]
Age-standardized	15.2 [11.5;20.3]	13.8 [11.7;16.4]	14.3 [11.9;17.1]	14.0 [11.7;16.7]	13.6 [12.1;15.4]
Colon–rectum					
15–44	55.2 [52.3;58.3]	55.1 [52.1;58.1]	55.2 [52.2;58.2]	55.1 [52.1;58.1]	55.9 [53.0;58.8]
45–54	54.6 [52.7;56.6]	54.4 [52.5;56.4]	54.5 [52.6;56.5]	54.4 [52.5;56.4]	55.0 [53.1;56.8]
55–64	52.2 [50.7;53.8]	51.9 [50.4;53.4]	52.5 [51.0;54.0]	52.1 [50.6;53.6]	53.1 [51.7;54.5]
65–74	47.0 [45.3;48.9]	47.3 [45.6;49.0]	47.5 [45.8;49.2]	46.9 [45.3;48.6]	49.5 [48.0;51.0]
75 and older	43.7 [37.5;51.0]	39.1 [35.6;42.9]	42.3 [38.5;46.4]	38.7 [35.2;42.5]	46.1 [44.7;47.5]
Nonstandardized	47.4 [44.9;49.9]	50.2 [49.2;51.3]	50.0 [48.9;51.0]	45.9 [45.0;46.9]	48.7 [48.0;49.4]
Age-standardized	48.7 [46.7;50.9]	47.4 [46.1;48.7]	48.5 [47.2;49.9]	47.2 [46.0;48.5]	50.4 [49.7;51.2]
Prostate					
15–54	60.5 [52.8;69.3]	60.3 [52.2;68.3]	60.2 [52.1;68.2]	60.5 [52.4;68.6]	61.4 [53.3;68.6]
55–64	73.7 [70.5;77.0]	71.9 [68.8;75.0]	73.9 [70.6;77.1]	73.8 [70.6;77.0]	75.8 [72.9;78.5]
65–74	68.6 [65.4;71.9]	67.8 [64.8;70.9]	69.8 [66.7;72.9]	69.1 [66.1;72.2]	73.0 [71.7;74.3]
75–84	48.3 [40.3;57.9]	49.8 [43.5;56.9]	55.4 [48.5;63.3]	51.7 [45.2;59.0]	60.7 [52.2;68.2]
85 and older	33.1 [11.6;94.7]	21.0 [6.1;71.9]	18.8 [5.5;64.3]	19.1 [5.6;65.6]	35.8 [0.0;98.6]
Nonstandardized	59.2 [53.1;66.1]	63.8 [61.7;65.9]	69.5 [67.2;71.8]	63.6 [61.5;65.7]	70.4 [68.6;72.2]
Age-standardized	61.4 [58.0;65.1]	60.4 [57.5;63.4]	62.5 [59.6;65.6]	61.5 [58.7;64.5]	66.4 [57.3;77.0]
Breast					
15–44	68.5 [67.1;70.0]	68.5 [67.0;69.9]	68.5 [67.1;69.9]	68.5 [67.1;69.9]	68.5 [67.1;69.8]
45–54	77.6 [76.5;78.8]	77.6 [76.5;78.7]	77.6 [76.5;78.7]	77.6 [76.5;78.7]	77.4 [76.4;78.5]
55–64	73.3 [71.9;74.6]	72.9 [71.5;74.2]	73.3 [71.9;74.6]	73.3 [71.9;74.6]	73.7 [72.5;74.9]
65–74	67.6 [65.6;69.8]	67.2 [65.1;69.2]	67.8 [65.8;69.8]	67.4 [65.4;69.5]	69.3 [67.4;71.1]
75 and older	47.1 [37.1;60.0]	49.6 [44.8;54.9]	53.8 [48.5;59.5]	50.5 [45.6;55.9]	56.3 [51.7;60.7]
Nonstandardized	67.6 [65.3;70.0]	71.6 [70.8;72.4]	71.7 [70.9;72.5]	69.6 [68.8;70.4]	72.5 [71.8;73.1]
Age-standardized	64.3 [61.0;67.7]	64.7 [63.1;66.4]	66.2 [64.5;68.0]	65.2 [63.5;66.8]	67.5 [66.0;68.9]
Thyroid gland					
15–44	98.7 [97.2;100]	98.6 [96.9;99.8]	98.6 [96.9;99.8]	98.6 [96.9;99.8]	99.3 [97.5;99.8]
45–54	93.3 [90.0;96.7]	93.4 [89.7;96.4]	93.4 [89.7;96.4]	93.3 [89.7;96.3]	93.9 [88.5;96.8]

Table 4. Estimates of 15-year net survival (age-standardized or not) obtained with Ederer I, Hakulinen, Ederer II, the UVM and the PPE (Continued)

	15-Year net survival				
Cancer site	Pohar-Perme	Ederer I	Hakulinen	Ederer II	Univariable model
55–64	93.2 [89.0;97.5]	93.5 [89.1;97.3]	94.2 [89.8;98.1]	93.2 [88.8;97.0]	95.1 [93.3;96.4]
65–74	75.0 [64.7;86.8]	75.6 [65.1;86.2]	76.7 [66.1;87.4]	74.6 [64.2;85.0]	77.6 [56.9;89.2]
75 and older	30.7 [11.5;82.1]	33.0 [13.4;76.0]	38.8 [15.8;89.5]	27.8 [11.3;64.2]	19.4 [14.2;25.2]
Nonstandardized	88.2 [85.1;91.4]	93.2 [91.1;95.2]	94.0 [91.9;96.0]	89.4 [87.4;91.3]	94.7 [91.2;96.8]
Age-standardized	82.3 [77.6;87.4]	82.9 [78.2;87.8]	84.0 [78.7;89.7]	81.9 [77.7;86.2]	82.0 [78.7;85.4]
Hodgkin disease					
15–44	88.1 [85.6;90.6]	88.0 [85.3;90.3]	88.1 [85.4;90.4]	88.0 [85.4;90.4]	88.1 [85.6;90.3]
45–54	70.9 [62.8;80.1]	71.1 [62.3;79.2]	71.2 [62.3;79.3]	70.8 [62.0;78.9]	69.9 [60.1;77.8]
55–64	62.3 [50.7;76.6]	62.6 [49.9;75.1]	63.7 [50.8;76.5]	62.1 [49.5;74.6]	64.1 [9.0;92.1]
65–74	35.9 [22.8;56.6]	35.5 [22.6;54.2]	36.9 [23.5;56.3]	27.1 [17.3;41.4]	38.9 [25.9;51.7]
75 and older[1]	–	–	–	–	–
Nonstandardized	73.5 [70.3;76.9]	80.6 [77.5;83.5]	79.9 [76.9;82.8]	73.9 [71.1;76.5]	79.3 [76.9;81.6]
Age-standardized[1]	–	–	–	–	–

[1]Not feasible.

large. Usually, age has an important impact on the ICM (thus, on the bias) because it has a great effect on the expected mortality and often a great effect on cancer mortality. This is reflected by the much more important magnitude of errors in the nonstandardized estimation *versus* age-class estimation or age-standardized estimation. However, the influences of the other life-table variables should not be ignored. The errors observed in age-class estimates may indeed be partially explained by the influences of these variables. Studying jointly the influence of all life-table variables is however complicated because this influence depends on the effects of *all* these variables together on cancer mortality *and* other-cause mortality as well as on the distribution of these variables.

A recent simulation[27] compared the following net-survival estimators: excess-rate regression models, Ederer I, Hakulinen and Ederer II methods. In this simulation, the ICM depended only on age. The authors concluded that the effect of age should be correctly taken into account to obtain unbiased estimations and that Ederer II was the less biased method among Ederer I, Hakulinen and Ederer II, which is in agreement with our present results. Hakulinen *et al.*[28] performed recently empirical comparisons between the Hakulinen and Ederer II methods. They concluded that Ederer II method should be favored for estimating net survival by cancer registries. Since then, findings by Pohar-Perme *et al.*[20] and Danieli *et al.*[21] clarified what these estimators actually estimate,[20] determined formally the sources of bias,[20] proposed a correct estimator,[20] and quantified the bias of the four classical methods on simulated data.[21] These points were not all fully considered in the publications of Rutherford *et al.*[27] or Hakulinen *et al.*[28]

One limitation of our study is that the difference between the estimates provided by any method M and PPE ones is only an approximation of the bias of M because of the uncertainty of the estimates with both methods. Another limitation is the length of the follow-up period; patients diagnosed after 1998 were followed-up for <10 years. Nonhomogeneous potential follow-ups (*e.g.*, dependent on age) may lead to bias in all estimators, including PPE.[20] However, Danieli *et al.*[21] have shown in a simulation study that even if the potential follow-ups were highly dependent on age, PPE performed much better than all the classical methods at 5 and 10 years postdiagnosis. The results at 15 years obtained in our study showed patterns that agreed with the simulation study of Danieli *et al*. These considerations are in favor of a much higher performance of PPE over the classical methods on long periods of follow-up (say, ≥10 years) with important administrative censorship. A third limitation is that some analyses at 15 years would have required larger datasets. It would also be interesting to study the bias over more than 15-year follow-up periods.

In the absence of the causes of death, net-survival estimators make the assumption that the available expected mortality rate reflects correctly the other-cause (than cancer) mortality rates. If not, this could induce bias in any estimator, including PPE.[29] It seems however that PPE is not more sensitive to the use of inappropriate expected mortality rates than other methods.[21]

Comparisons of cancer survivals between different populations (*e.g.*, from different regions or diagnosed at different periods) is a major epidemiological objective. It requires removing the effect of other-cause mortality[1] which may be different between populations. This was at the origin of the concept of net survival. Ederer I, Hakulinen, Ederer II and the UVM fail to estimate correctly net survival because they

fail to remove completely the other-cause mortality.[20] Thus; when the net survivals in two countries are the *same*, the estimates obtained with any classical method (Ederer II, for example) will be *different* if the other-cause mortalities are different. International comparisons will be therefore misleading even if they use the same method, say Ederer II.

Using PPE, our team was able to obtain estimations of the net survival by sex, age, and period of diagnosis as well as age-standardized net survivals by sex and period of diagnosis on 55 cancer sites using data of Francim network.[22,23] Technical feasibility is an important feature of such studies. PPE is implemented in the *relsurv* package on R software and we have adapted this code to deal with large datasets. Nevertheless, PPE has some drawbacks, mainly those of a nonparametric estimator. The survival curves obtained with PPE may be erratic, which might be difficult to understand by nonspecialists and the variance of PPE may be larger than the one obtained through a multivariable excess-hazard modeling approach. However, Danieli *et al.*[21] have shown that the PPE had a high performance in terms of bias-variance tradeoff: PPE showed a much higher performance than those of the classical methods, and a slightly lower performance than that of a multivariable model. We have also explored the building of a multivariable model able to take into account the ICM by including all life-table variables.[20,21] As mentioned previously, we found no single strategy suitable to all cancer sites. Lately, net survival was estimated on English cancer registry data by a flexible multivariable model;[30] however, full details on that model are not yet available.

In conclusion, it is technically feasible to use the PPE in survival studies that use registry data, and we strongly recommend the PPE for routine point estimations of net survivals by cancer registries. We see no reason to favor any classically used method such as Ederer I, Hakulinen, Ederer II or the univariable excess-rate regression models because, unlike the PPE, they are all biased.

Acknowledgements

The research was carried out within the context of a four-institute research-program partnership that involved the Institut National du Cancer (INCa), the Institut de Veille Sanitaire (InVS), FRANCIM and Hospices Civils de Lyon.

References

1. Esteve J, Benhamou E, Raymond L. Statistical methods in cancer research, vol. 4. Descriptive epidemiology. Lyon: IARC Scientific Publications, 1994. 302 p.
2. Percy C, Stanek E 3rd, Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. *Am J Public Health* 1981;71: 242–50.
3. Ashworth TG. Inadequacy of death certification: proposal for change. *J Clin Pathol* 1991;44:265–8.
4. Esteve J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. *Stat Med* 1990;9:529–38.
5. Remontet L, Bossard N, Belot A, Esteve J. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Stat Med* 2007;26:2214–28.
6. Rachet B, Woods LM, Mitry E, et al. Cancer survival in England and Wales at the end of the 20th century. *Br J Cancer* 2008;99 (Suppl 1): S2–10.
7. Bossard N, Velten M, Remontet L, et al. Survival of cancer patients in France: a population–based study from The Association of the French Cancer Registries (FRANCIM). *Eur J Cancer* 2007;43: 149–60.
8. Chirlaque MD, Salmeron D, Ardanaz E, et al. Cancer survival in Spain: estimate for nine major cancers. *Ann Oncol* 2010;21 (Suppl 3):iii21–iii29.
9. Coleman MP, Forman D, Bryant H, et al. Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995–2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data. *Lancet* 2011;377:127–38.
10. Ederer F, Axtell LM, Cutler SJ. The relative survival rate: a statistical methodology. *Natl Cancer Inst Monogr* 1961;6:101–21.
11. Hakulinen T. Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics* 1982;38:933–42.
12. Altekruse SF, Kosary CL, Krapcho M, et al. SEER cancer statistics review, 1975–2007. Bethesda, MD: National Cancer Institute, 2010. Available at: http://seer.cancer.gov/csr/1975_2007/. Accessed March 30, 2012.
13. Howlader N, Noone AM, Krapcho M, et al. SEER Cancer Statistics Review, 1975–2008. Bethesda, MD: National Cancer Institute, 2011. Available at: http://seer.cancer.gov/csr/1975_2008/. Accessed March 30, 2012.
14. de Angelis R, Francisci S, Baili P, et al. The EUROCARE-4 database on cancer survival in Europe: data standardisation, quality control and methods of statistical analysis. *Eur J Cancer* 2009; 45:909–30.
15. Coleman MP, Quaresma M, Berrino F, et al. Cancer survival in five continents: a worldwide population-based study (CONCORD). *Lancet Oncol* 2008;9:730–56.
16. Cancer Registry of Norway. Cancer in Norway 2009—cancer incidence, mortality, survival and prevalence in Norway. Oslo: Cancer Registry of Norway, 2011. 169 p.
17. Finnish Cancer Registry. Cancer in Finland 2006 and 2007. Helsinki: Cancer Society of Finland Publication, 2009. 87 p.
18. Engholm G, Gislum M, Bray F, Hakulinen T. Trends in the survival of patients diagnosed with cancer in the Nordic countries 1964–2003 followed up to the end of 2006. Material and methods. *Acta Oncol* 2010;49:545–60.
19. Hakulinen T. On long-term relative survival rates. *J Chronic Dis* 1977;30:431–43.
20. Perme MP, Stare J, Esteve J. On estimation in relative survival. *Biometrics* 2012;68:113–20.
21. Danieli C, Remontet L, Bossard N, et al. Estimating net survival: the importance of allowing for informative censoring. *Stat Med* 2012;31:775–86.
22. Monnereau A, Troussard X, Belot A, et al. Unbiased estimates of long-term net survival of haematological malignancy patients detailed by major subtypes in France. *Int J Cancer*, in press.
23. Jooste V, Grosclaude P, Remontet L, et al. Unbiased estimates of long-term net survival of solid cancers in France. *Int J Cancer*, in press.
24. Wood SN. Generalized additive models: an introduction with R. Boca Raton: Chapman & Hall/CRC, 2006. 410 p.
25. Corazziari I, Quinn M, Capocaccia R. Standard cancer patient population for age standardising survival ratios. *Eur J Cancer* 2004;40:2307–16.
26. Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In: Jewel N, Dietz K, Farewell V, eds. Aids epidemiology: methodological issues. Boston, MA: Birkhäuser, 1992. 297–331.
27. Rutherford MJ, Dickman PW, Lambert PC. Comparison of methods for calculating relative survival in population-based studies. *Cancer Epidemiol* 2012;36:16–21.
28. Hakulinen T, Seppa K, Lambert PC. Choosing the relative survival method for cancer survival estimation. *Eur J Cancer* 2011;47:2202–10.
29. Sarfati D, Blakely T, Pearce N. Measuring cancer survival in populations: relative survival vs. cancer-specific survival. *Int J Epidemiol* 2010;39: 598–610.
30. Office for National Statistics. Statistical Bulletin. Cancer Survival by Cancer Network in England—patients diagnosed 1996–2009 and followed up to 2010. Available at: http://www.ons.gov.uk/ons/rel/cancer-unit/cancer-survival-by-cancer-network/patients-diagnosed-in-1996-2009-followed-up-to-2010/cancer-survival-by-cancer-network-sb2.html. Accessed March 30, 2012.

II.2.6.b. Réactions suscitées

Le sujet survie nette/survie relative étant d'actualité, cet article a fait réagir la communauté scientifique et a suscité de nombreux débats [Dickman, 2013][Roche, 2013].

Jusqu'à récemment, les méthodes de survie relative étaient privilégiées dans certaines études de survie : la SEER a pendant longtemps utilisé la méthode d'Ederer I avant de passer à la méthode d'Ederer II en 2011, Eurocare a toujours utilisé la méthode de Hakulinen. Cependant, comme nous l'avons évoqué précédemment, ces méthodes n'estiment pas la survie nette. Le but de cet article est donc de communiquer avec les cliniciens et les épidémiologistes sur les avancées qu'il y a eu ces dernières années sur le plan méthodologique de l'estimation de la survie nette à travers un article épidémiologique, qui sera peut-être plus pertinent pour attirer leur attention. Ils pourront donc se rendre compte de l'erreur causée par l'utilisation de méthodes biaisées pour l'estimation de la survie nette.

Cependant, cet article a reçu quelques critiques dans le fait de promouvoir la méthode de Pohar-Perme au détriment des autres méthodes existantes [Dickman, 2013]. Il a été notamment soulevé le fait que l'ampleur des erreurs pour certaines méthodes, particulièrement la méthode d'Ederer II, est assez faible et que la méthode ne mérite donc pas d'être exclue des méthodes à utiliser ; ou encore que la méthode de Pohar-Perme, méthode non-paramétrique, possède une variabilité importante, surtout à long terme chez les personnes âgées. Ces différentes remarques montrent la difficulté du message à faire passer. En effet, que l'erreur soit importante ou pas, si l'estimateur estime théoriquement une autre quantité que la survie nette, il ne doit pas être utilisé pour estimer cette survie nette. De même, le fait d'observer de la variabilité dans les courbes de survie à long terme obtenue avec la méthode de Pohar-Perme ne veut pas dire que cet estimateur est « instable » : cela traduit simplement qu'il y a peu d'informations dans les données pour estimer la survie nette. Ce manque d'information est particulièrement présent à long terme chez les personnes âgées pour lesquelles les risques de décès dus aux autres causes sont très élevés. Par ailleurs, l'étude de simulation a pu montrer que cette méthode était la plus performante des méthodes non-paramétriques en terme de compromis biais-variance (critère RMSE), même à long terme (à 10 ans) [Danieli, 2012].

Cette étude a permis d'illustrer sur données réelles les erreurs que l'on pourrait obtenir en utilisant une autre méthode que celle de Pohar-Perme. Il s'agit en aucun cas d'une étude visant à comparer les différentes méthodes du fait que sur données réelles, la survie nette

théorique n'est pas connue et des critères de performances tels que le biais, l'erreur quadratique moyenne et le taux de couverture ne peuvent pas être estimés. Les trois articles, théorique [Pohar-Perme, 2012], méthodologique [Danieli, 2012] et épidémiologique [Roche, 2013] sont tous complémentaires les uns des autres, ils doivent être tous considérés afin d'éviter des incompréhensions telles qu'on a pu le voir dans la critique [Dickman, 2013].

IJC
International Journal of Cancer

Estimating net survival in population-based cancer studies

Paul W. Dickman[1], Paul C. Lambert[1,2], Enzo Coviello[3] and Mark J. Rutherford[2]

[1] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[2] Department of Health Sciences, Centre for Biostatistics and Genetic Epidemiology, University of Leicester, United Kingdom
[3] Epidemiology Unit, Cancer Registry ASL BT, Barletta, Italy

Roche *et al.*[1] conclude that "great errors" occur in "all classical methods" of estimating net survival and suggest that "cancer registries should abandon all classical methods and adopt the new Pohar Perme [PP] method." The PP method may well be technically superior, but we believe Roche *et al.* misrepresent the bias in the preferred classical methods in a manner that may unduly cause alarm. It might be some time before the PP method is implemented in the software used by many registries and it would be a great pity if the scientific community dismissed methodologically sound and important research because of Roche and coworker's claims of "great errors."

It is no surprise that the unstandardized "classical methods" reported by Roche *et al.* are biased.[2–5] However, most cancer registries report age-specific or age-standardized estimates of survival. The differences between the age-standardized PP and the corresponding Ederer II estimates (the preferred classical estimator[2]) were small (< 0.5 and < 1 percent units at 5 and 10 years respectively). Roche *et al.* compare the PP method to four "classical methods" including one (Ederer I) that has not been widely used for 50 years, another (Hakulinen) that the original author no longer recommends,[2] and one (UVM) that shows greater discrepancy than usually observed (see our later comments). They apply the five methods to several cancer sites and selectively cite, *e.g.*, in the abstract, the largest observed differences (ignoring random error and multiple testing) to support their claims that "great errors" occur in "all classical methods." We do not dispute that the PP method is the only unbiased estimator of net survival. Our criticism is that the comparisons by Roche *et al.* are not objective; they grossly overstate the magnitude of the bias in the Ederer II method in a manner that could mislead and alarm the research community.

An important issue with the PP estimator is the increased variability and lack of stability for long-term survival, particularly for older age groups. This is illustrated in Figure 1 where the PP estimates are shown together with estimates from a univariable model. Our univariable model is a flexible parametric model[6] and is conceptually similar but technically different to that used by Roche *et al.*; we fitted the model separately to each age group and used restricted cubic splines to model the effect of time with no additional covariates. The model-based approach provides greater stability and less variability, *i.e.*, narrower confidence intervals, for long-term estimates.

With respect to Figure 1 there are two statements that are undeniably true—the PP method provides unbiased estimates of the true net survival and the true net survival is unknown. The bias in the "classical methods" results from a dependence between cancer mortality and non-cancer mortality. All approaches are unbiased when applied to patients with the same expected survival, *e.g.*, of the same sex, age and year of diagnosis. The bias in the classical methods is smaller when they are applied within narrow age groups, as is done when age standardizing. The bias is largest in the oldest age group, where expected survival is most heterogeneous, but this bias can be reduced by adjusting for age.

Unbiasedness is undoubtedly a desirable property in an estimator, but does not mean that all estimates will equal the true value. Estimates are subject to random variation and it so happens that the PP estimator has larger variability than other methods.[3] The approach used by Roche *et al.* to calculate the "bias with the classical methods" is fundamentally flawed. They calculate the difference between the PP estimate and the classical estimate (both estimated with error) and call this difference the "bias" or "error." Bias is defined as the difference between the expected value of an estimator and the true value, not the difference between two estimates. Roche *et al.* mention this in the discussion, but appear to ignore the implications when reporting results and making very strong recommendations. Although we cannot guarantee the model-based estimates in Figure 1 are unbiased, we would argue that any bias is small. The two estimates presented in Figure 1 differ in some instances by more than 15 percent units. These differences do not represent the bias in the model-based estimates as Roche *et al.* would have us believe; they instead represent the large variability.

The large variability in long-term survival, particularly for older age groups, is inherent in the data. The PP estimator weights each observation by the inverse of the expected survival, providing a very elegant approach to estimating the true net survival for a cohort of individuals with heterogeneous expected survival times. Relatively few patients aged 75 years or more at diagnosis survive 10 years or more, and those who do will contribute large weight, potentially causing large fluctuations when these individuals die. The model-based approach has lower variance because it relies on assumptions not required by the PP method. We believe these assumptions are appropriate and the resulting stability and lower variance in estimates of long-term survival are desirable properties. We recognize that we focus heavily on methods for estimating

Int. J. Cancer: 133, 519–522 (2013)

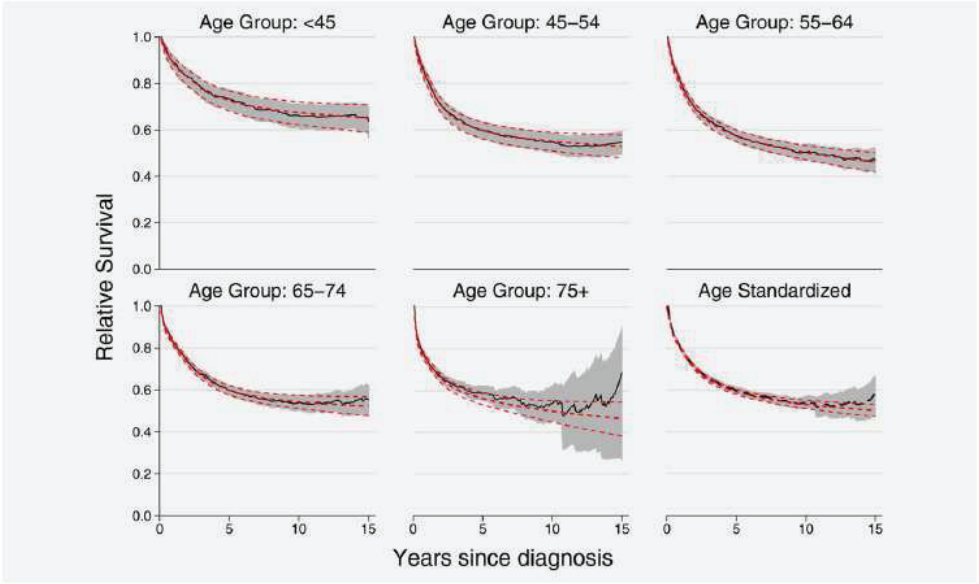


Figure 1. Estimates of net survival for men diagnosed with colon cancer in Finland 1992–2007. The solid black lines show the Pohar Perme estimates and the gray area the corresponding 95% pointwise confidence limits. The model-based estimates are shown using dashed red lines along with the corresponding 95% pointwise confidence limits. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

long-term survival for the oldest age group, which is of lesser practical interest given the paucity of available data. Our focus on this area is a response to Roche *et al.*'s choice to cite, in their abstract, results from long-term estimates in a way which we believe misrepresents the truth.

The traditional method of age standardization involves first estimating net survival for each of several age groups and then averaging these estimates using weights from the standard population. If age-specific estimates are obtained using the PP method there may be large variability for longer-term survival for older age groups (Fig. 1). These older age groups have relatively large weight in many standard populations (*e.g.*, 58% over the age of 65[7]), which leads to age standardized estimates with large variance (Fig. 1).

We are surprised by the large "bias" in the univariable model reported by Roche *et al.* We have never observed differences in age-specific estimates estimators of the magnitude they report. The model we use has restrictions to ensure that it is not overly influenced by sparse data in the tails.

Roche *et al.* claim to "see no reason to favor any classically used method such as Ederer I, Hakulinen, Ederer II." One additional reason is that the PP estimator is not available in software commonly used by cancer registries, such as SEER*Stat. We do not believe that immediately abandoning existing software is motivated given the magnitude of the bias for the recommended "classical method" (Ederer II) is not large for many of the analyses most commonly performed by cancer registries. We are pleased to have incorporated the PP estimator into our own software[8] and hope the developers of SEER*Stat incorporate it into their software so it may be widely used. We feel, however, that the best of the classical methods are not as inferior as Roche *et al.* would have us believe. Researchers should also be aware that the lack of bias in the PP estimator comes at a price of higher variance and this becomes an important issue when age-standardization is employed or with small data and/or long follow-up.

Paul W. Dickman[1]
Paul C. Lambert[1,2]
Enzo Coviello[3]
Mark J. Rutherford[2]

References

1. Roche L, Danieli C, Belot A, et al. Cancer net survival on registry data: Use of the new unbiased Pohar-Perme estimator and magnitude of the bias with the classical methods. Int J Cancer, in press.
2. Hakulinen T, Seppä K, Lambert PC. Choosing the relative survival method for cancer survival estimation. *Eur J of Cancer* 2011;47:2202–10.
3. Pohar Perme M, Stare J, Estève J. On estimation in relative survival. *Biometrics* 2012;68:113–20.
4. Rutherford MJ, Dickman PW, Lambert PC. Comparison of methods for calculating relative survival in population-based studies. *Cancer Epidemiol* 2012;36:16–21.

5. Danieli C, Remontet L, Bossard N, et al. Estimating net survival: the importance of allowing for informative censoring. *Stat Med* 2012;31:775–86.
6. Royston P, Lambert PC. Flexible parametric survival analysis in Stata: Beyond the Cox model. Stata Press, College Station, 2011.
7. Corazziari I, Quinn M, Capocaccia R. Standard cancer patient population for age standardising survival ratios. *Eur J Cancer* 2004;40:2307–16.
8. Dickman PW, Coviello E, Hills M. Estimating and modelling relative survival using Stata, 2012. http://pauldickman.com/rsmodel/stata_colon/. Last accessed February 3, 2013.

DOI: 10.1002/ijc.28041
History: Received 20 Oct 2012; Accepted 29 Nov 2012; Online 22 Jan 2013

Correspondence to: Paul W. Dickman, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Box 281, 171 77 Stockholm, Sweden, Tel.: [+46-8-5248–6186], Fax: [+46-8-314–975], E-mail: paul.dickman@ki.se

IJC
International Journal of Cancer

Author's reply to: Estimating net survival in population-based cancer studies

Laurent Roche[1,2,3,4], Coraline Danieli[1,2,3,4], Aurélien Belot[1,2,3,4,5], Jean Iwaz[1,2,3,4], Laurent Remontet[1,2,3,4] and Nadine Bossard[1,2,3,4]

[1] Hospices Civils de Lyon, Service de Biostatistique, F-69003, Lyon, France
[2] Université de Lyon, F-69000, Lyon, France
[3] Université Lyon 1, F-69100, Villeurbanne, France
[4] CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biostatistique-Santé, F-69100, Villeurbanne, France
[5] Institut de Veille Sanitaire, Département des maladies chroniques et traumatismes, F-94410, Saint-Maurice, France

Letter to the Editor

Dear Editor,

We were surprised by the hostile tone of the criticism of our work[1] made by Dickman *et al.* (Dickman PW, Lambert PC, Coviello E, et al. Estimating net survival in population-based cancer registries, Submitted to *Int J Cancer* 2012). Nevertheless, we develop herein our response to this stern criticism.

Over decades, a great confusion existed between "relative survival" and "net survival" (NS).[2] Recently, Pohar-Perme *et al.*[3] clarified completely this issue and showed theoretically that: (*i*) the "classical" methods do not estimate NS; (*ii*) the nonparametric Pohar-Perme estimator[3] (PPE) is an unbiased estimator of NS. In 2012, the higher performance of PPE over the classical methods was assessed by Danieli *et al.*[4]

Thus, our work had not to prove once more the superiority of PPE: our recommendation of PPE for routine estimation of NS was based on previous works.[3,4] Convinced that the work of Pohar-Perme *et al* is a major advance in NS estimation, our exact aim was to share this advance with all epidemiologists and cancer registry personnel and show the differences that may be expected between classical methods and PPE on real data.

We computed the differences between the estimates provided by each classical method *vs.* PPE as reference. These differences were called "errors" and these errors are measures of bias that have, obviously, certain variabilities. In fact, Dickman *et al.* thought that our aim was to prove the superiority of PPE but this was not our aim; thus, the claims that our "comparisons are not objective" and that our approach is "fundamentally flawed" are off-topic.

Nevertheless, Dickman's comments call for several clarifications, especially regarding variance and variability.

Firstly, our work discussed about the variability of PPE by citing the work of Danieli *et al.*[4] which is, thus far, the single work about the variability of PPE. Besides, this work underlined the superiority of PPE over all classical methods in terms of bias-variance tradeoff, which was ignored by Dickman *et al.* What does then the "price of higher variance" mean here?

Secondly, the variance of PPE may be large for long-term NS estimations in elderly groups; however, this is a feature of the data and not an undesirable property of PPE. Indeed, PPE is the only non-parametric unbiased estimator of NS, and, as such, the only one that reflects correctly the variability due to lack of data. Thus, interestingly, the PPE survival curve relative to the elderly group in Figure 1 by Dickman *et al.* suggests strongly very scarce information after 10 years. In such a case, model-based approaches (*e.g.*, the univariable model) rely on unassessable assumptions because of lack of data. Besides, a narrow calculated confidence interval is no proof of high performance of an estimator; the coverage rate of this confidence interval should be examined too. Beyond this, the concept of NS may be questioned in long-term survival of elderly groups; crude survival would be more relevant despite its dependence on other-cause mortality.

Thirdly, claiming that cancer registries do not report nonstandardized estimates would be untrue (most cancer registries do) and the criticisms about the choice of the methods we compared are based on wrong arguments. Until recently, the main publications on cancer survival used Ederer I and Hakulinen (*e.g.*, SEER switched from Ederer I to Ederer II in 2011[5] following a presentation by Estève and Hakulinen was used in Eurocare 4[6]). Besides, our results with the univariable model agree with previous ones[4] and we did not succeed in identifying the "doubts" by Dickman *et al.* Furthermore, this univariable model was used previously (France in 2007[7]).

Finally, supporting Ederer II on the basis of empirical errors is not scientifically relevant; the correct solution (PPE) is now available. Furthermore, we do not agree that deciding to use a method or not depends on its availability in a specific software. PPE is already available in R software (package *relsurv*) and a new *stns* command will soon be released in Stata (Clerc-Urmès I, Grzebyk M, Hédelin G. Net survival estimation with stns. Submitted to *Stata Journal* 2012) (available from michel.grzebyk@inrs.fr).

Dickman *et al.* recognize that PPE is unbiased and hope it will "be widely used" despite a "price." They then propose a univariable approach but admit a multivariable approach would reduce the bias. They also still support the use of

Ederer II. It is in fine unclear which method they would finally recommend to cancer registries.

Yours sincerely,
On behalf of the authors
Laurent Roche

References

1. Roche L, Danieli C, Belot A, et al. Cancer net survival on registry data: use of the new unbiased Pohar-Perme estimator and magnitude of the bias with the classical methods. *Int J Cancer* 2012, doi: 10.1002/ijc.27830. [Epub ahead of print].
2. Esteve J, Benhamou E, Croasdale M, et al. Relative survival and the estimation of net survival: elements for further discussion. *Stat Med* 1990;9:529–38.
3. Pohar-Perme M, Stare J, Esteve J. On estimation in relative survival. *Biometrics* 2012;68:113–20.
4. Danieli C, Remontet L, Bossard N, et al. Estimating net survival: the importance of allowing for informative censoring. *Stat Med* 2012; 31:775–86.
5. Cho H, Howlader N, Mariotto AB, et al. Estimating relative survival for cancer patients from the SEER Program using expected rates based on Ederer I versus Ederer II method. Surveillance Research Program, National Cancer Institute; 2011. Technical Report no. 2011-01. Available from: http://surveillance.cancer.gov/reports/. Accessed February 4, 2013.
6. De Angelis R, Francisci S, Baili P, et al. The EUROCARE-4 database on cancer survival in Europe: data standardisation, quality control and methods of statistical analysis. *Eur J Cancer* 2009;45:909–30.
7. Bossard N, Velten M, Remontet L, et al. Survival of cancer patients in France: a population–based study from The Association of the French Cancer Registries (FRANCIM). *Eur J Cancer* 2007;43:149–60.

DOI: 10.1002/ijc.28039

History: Received 26 Nov 2012; Accepted 29 Nov 2012; Online 22 Jan 2013

Correspondence to: Laurent Roche, Service de Biostatistique des Hospices Civils de Lyon, 162 avenue Lacassagne, F-69424 Lyon Cedex 03, France, Tel.: [(33)−4–72-11–57-55], E-mail: laurent.roche01@chu-lyon.fr

Chapitre III

Tests des hypothèses des modèles paramétriques du taux en excès

III.1 Introduction

Le modèle paramétrique du taux de mortalité en excès est utilisé dans le but de donner des estimations ponctuelles de la survie nette des patients atteints de cancer mais également dans le but d'identifier les principaux facteurs pronostiques influençant le taux de mortalité en excès, ceci n'étant pas possible avec la méthode non-paramétrique de Pohar-Perme. Nous avons vu précédemment que cette modélisation devait pouvoir prendre en compte l'ensemble des variables introduisant une censure informative (celle définissant la table de mortalité). Cependant, la spécification du modèle multivarié nécessite de faire des hypothèses afin de modéliser au mieux l'effet des covariables d'intérêt et, en finalité, la survie lorsque cela est l'objectif. Les principaux points nécessitant des hypothèses sont les suivants : (i) A quelle question voulons-nous répondre et par conséquent quelles covariables doivent être incluses dans le modèle ? (iii) Quelle forme donner au taux de mortalité de base ? (iv) Comment modéliser les effets des covariable ? constants au cours du temps ou dépendants du temps ? Si des effets sont dépendants du temps, doit-on utiliser une fonction linéaire au cours du temps ou une fonction plus flexible telle qu'un spline pour modéliser cette dépendance temporelle ? Si la covariable est continue, son effet est-il linéaire ou non ? (v) Quelle fonction de lien utiliser ? (vi) Comment prendre en compte les interactions entre covariables ? Les décisions prises sont importantes car un modèle inapproprié peut conduire à des conclusions erronées. Cette stratégie devra également permettre de fournir un modèle prédictif, robuste et parcimonieux. Le modèle retenu est généralement celui possédant le meilleur compromis entre complexité (parcimonie), biais et variance.

Lorsqu'une hypothèse est émise au sujet d'une covariable, l'étape suivante consiste à évaluer si les données ne contredisent pas cette hypothèse. La vérification des hypothèses, qui

peut être effectuée à l'aide d'un test formel ou graphiquement, repose principalement sur les résidus. Les résidus représentent, d'une façon générale, la différence entre ce qui est observé et ce qui est prédit par le modèle. La difficulté dans l'interprétation des résultats graphiques que fournissent certains résidus est que l'on ne sait pas toujours évaluer l'adéquation du modèle aux données du fait que l'on ne sait pas si la tendance obtenue est due à une mauvaise spécification du modèle ou à une variation aléatoire. Une solution judicieuse est alors de travailler avec des résidus dont la distribution sous l'hypothèse nulle H_0 est connue (H_0 : « l'hypothèse est correcte »), ce qui permet d'élaborer un test formel ; ce type de résidus s'appuie généralement sur les processus stochastiques.

En effet, les résidus existants ayant eu le plus de succès [Barlow, 1988][Therneau, 1990] sont ceux qui ont été développés dans le cadre de la théorie des processus de comptage. Pour un individu *i*, au lieu de considérer la variable aléatoire T_i, comme cela a été fait précédemment, la survenue de l'évènement sera décrit par le processus ponctuel associé N_i. L'évènement considéré dans notre cas étant le décès, N_i est une fonction en escalier qui vaut 0 tant que l'évènement n'a pas eu lieu et 1 dès que le patient décède. Si le patient est censuré, N_i est nul tout au long du suivi.

$$N_i(t) = 1\{T_i \le t, \delta_i = 1\}$$

De la même manière, la présence du patient *i* dans la population au temps *t* est décrite par le processus Y_i qui vaut 1 tant que le patient est toujours à risque dans la population et 0 s'il n'est plus à risque (censuré, perdu de vue, ou décédé).

$$Y_i(t) = 1\{T_i \ge t\}$$

En supposant que l'on ait une population de *n* patients, le processus de comptage *N* et le processus d'indicateur de présence *Y* sont les suivants :

$$N(t) = \sum_{i=1}^{n} N_i(t)$$

$$Y(t) = \sum_{i=1}^{n} Y_i(t)$$

La théorie des processus de comptage repose principalement sur la décomposition de Doob-Meyer [Meyer, 1962][Meyer, 1963] qui dit qu'en survie, en raisonnant à l'échelle individuelle, le processus de comptage $N_i(t)$ pour le patient i faisant partie d'une population de taille n, peut s'écrire comme la somme d'une martingale $M_i(t)$ et d'un processus prévisible croissant $\Lambda_i(t)$ tel que $E(\Lambda_i(t)) < +\infty$ pour tout t et $\Lambda_i(0) = 0$. Cette décomposition est unique :

$$N_i(t) = \Lambda_i(t) + M_i(t)$$
$$N_i(t) = \int_0^t Y_i(u)\lambda_i(u)du + M_i(t)$$

Cette décomposition peut être vue comme la décomposition classique « données = partie attendue + aléa », ou encore de manière analogique comme « Y = βX + ε » avec Y, X et ε représentant respectivement la variable expliquée, les variables explicatives et le bruit.

La martingale $M_i(t)$ est alors assimilé à un résidu. Elle a la propriété suivante :

$$E(M_i(t)) = 0 \ \forall i$$

En nous plaçant dans la situation où le taux de mortalité du patient i prend la forme usuelle :

$$\lambda_i(t) = \lambda_0(t)\exp(\beta z_i) \qquad \text{(III.1)}$$

avec $\lambda_0(t)$ représentant le taux de base au temps t, z_i représentant le vecteur des covariables incluses dans le modèle pour le patient i et β représentant le vecteur de leurs effets, le processus des résidus de martingale individuel [Therneau, 1990] s'écrit de la manière suivante :

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(u)\hat{\lambda}_0(u)\exp\left(\hat{\beta}z_i\right)du \qquad \text{(III.2)}$$

Il représente pour l'individu i, la différence entre le processus de comptage observé au temps t (binaire 0/1), et le compensateur de ce processus estimé par le modèle au temps t. D'une

façon plus générale, cela représente, pour le patient *i*, la différence entre le nombre d'évènements observés au temps *t* et celui qui serait attendu par le modèle au temps *t*.
Que l'on se place dans un modèle semi-paramétrique (modèle de Cox) ou un modèle paramétrique, les processus des résidus de martingale s'expriment de cette même façon (avec $\lambda_0(u)du$ estimé à l'aide de l'estimateur de Breslow dans le cas du modèle de Cox).

III.2 Objectif

Très peu d'outils diagnostics existent en survie nette. L'objectif principal de ce travail est donc de développer une « boîte à outils » composée de différents tests formels permettant de vérifier certaines hypothèses d'un modèle paramétrique du taux en excès et d'en évaluer les performances. Les hypothèses d'intérêt qui seront étudiées concernent l'hypothèse de proportionnalité des taux, la forme fonctionnelle et la fonction de lien. Nous nous efforcerons de développer ces tests dans un même cadre théorique. Pour cela, une étude des principaux résidus existant dans le cadre de la survie globale a été effectuée au préalable, afin de s'approprier les outils, de comprendre leur principe et de pouvoir les adapter en survie nette.

III.3 Résidus existants développés dans le cadre de la survie globale

Dans cette partie sont présentés tout d'abord les résidus les plus pertinents développés dans le cadre de la survie globale tels que les résidus de martingale [Therneau, 1990] et les résidus de Schoenfeld [Schoenfeld, 1982] permettant de vérifier respectivement la forme fonctionnelle et l'hypothèse des taux proportionnels. Sont ensuite présentés les résidus développés dans le cadre des processus de comptage qui permettent de vérifier les mêmes hypothèses ainsi que la forme de la fonction de lien. Le travail effectué dans cette partie a été de comprendre intuitivement, théoriquement et méthodologiquement chacun de ces résidus, en se focalisant principalement sur ceux développés dans le cadre des processus de comptage, afin de pouvoir les adapter dans le cadre de la survie nette.

III.3.1 Résidus de martingale

Définition

Les résidus de martingale [Therneau, 1990] représentent pour l'individu *i*, la différence entre l'évènement observé (binaire 0/1), et le taux de mortalité prédit par le modèle au temps T_i, temps de suivi observé pour le patient *i*. D'une façon plus générale, cela représente, pour le patient *i*, la différence entre l'évènement observé et ce qui serait attendu par le modèle au temps T_i (correspond au processus des résidus de martingale individuel (III.2) estimé au temps T_i) :

$$\hat{M}_i(T_i) = N_i(T_i) - \int_0^{T_i} \hat{\lambda}_0(u)\exp\left(\hat{\beta}z_i\right)du \qquad \text{(III.3)}$$

Ces résidus sont asymétriques et négatifs pour les patients censurés.

En utilisant la décomposition de Doob-Meyer pour le processus N_i :

$$M_i(T_i) = N_i(T_i) - \int_0^{T_i} \lambda_i(u)du \text{ , avec notamment } \forall i \text{ , } E\left(M_i(T_i)\right) = 0$$

Pour que $\int_0^{T_i} \hat{\lambda}_0(u)\exp\left(\hat{\beta}z_i\right)du$ estime correctement $\int_0^{T_i} \lambda_i(u)du$, i.e. le modèle est correct, cela implique que :

$$E\left(\hat{M}_i(T_i)\right) \approx 0$$

Principe

Afin d'illustrer en quoi ces quantités sont des résidus, nous allons les utiliser sur données simulées afin d'en comprendre le principe. Dans cet exemple, la covariable d'intérêt est l'âge. Cette covariable suit une loi uniforme qui génère 26.7% de patients compris dans la classe d'âge [30,65], 33.3% de patients compris dans la classe d'âge [65,75] et 40% de patients compris dans la classe d'âge [75,85]. Cette covariable aura un effet linéaire-proportionnel sur le taux de mortalité qui s'exprimera alors de la manière suivante :

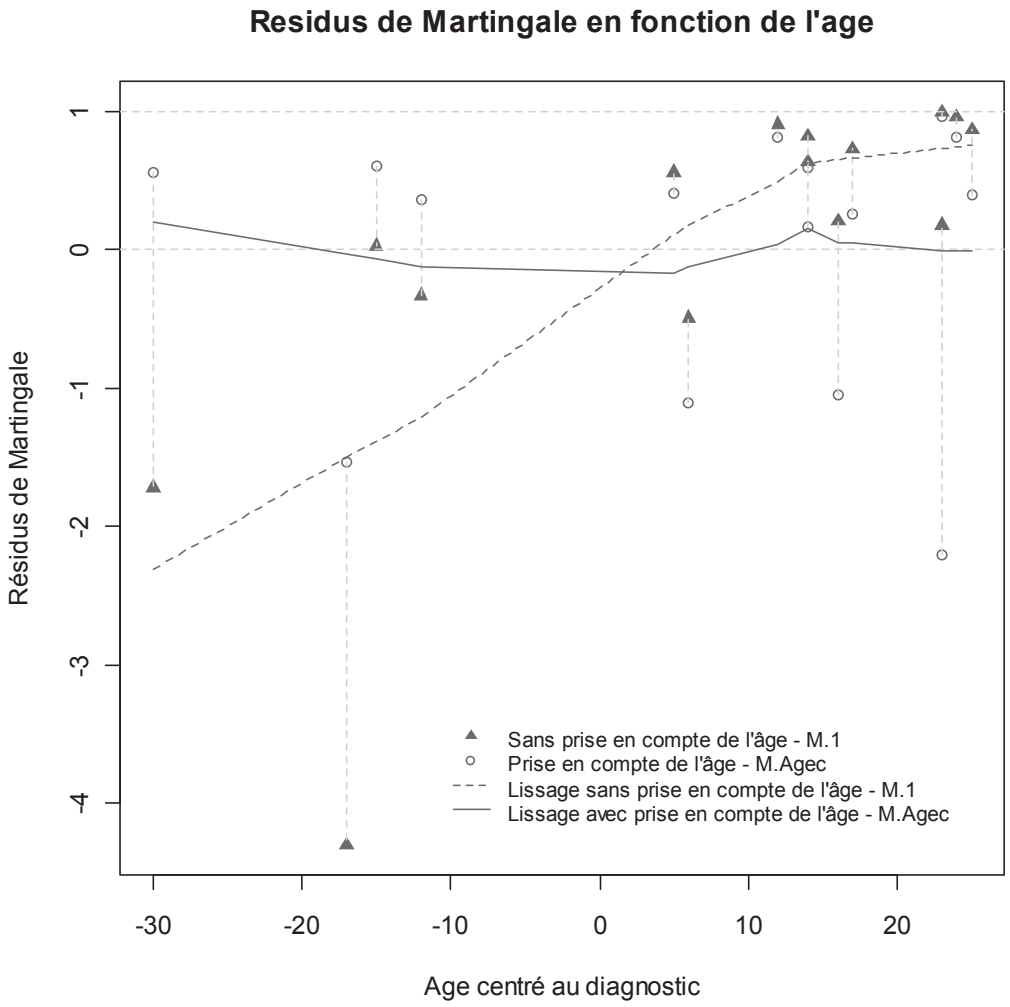
$\lambda(t) = \lambda_0(t) \times \exp(\beta_{Age} \times Agec)$ avec $\beta_{Age} = 0.05$. La covariable *année de diagnostic* suit une loi uniforme entre 1980 et 1985. Le temps de survie des patients *T* a été généré à l'aide de la méthode de la transformation inverse à partir d'une distribution exponentielle avec le paramètre $\lambda_0(t) = \lambda_0 = 0.5$. La censure administrative a été fixée à 1995. Le temps potentiel de suivi T_{max} est donc égale à 1995 moins l'année de diagnostic. Le temps final d'observation est donc égale au minimum de $\{T, T_{max}\}$. Les données sont alors composées d'une variable *Etat* indiquant si le patient est décédé pendant le temps d'observation ou s'il a été censuré. Un échantillon constitué de 15 patients a été généré.

Les données générées sont les suivantes :

Table III.1. Description des données simulées suivant le modèle $\lambda(t) = 0.5 \times \exp(0.05 \times Agec)$

Patient	Age	Agec	DureeDeVie	Etat	Rm.1	Rm.Agec
1	*42.99925*	*-17*	*10.0000000*	*0*	*-4.29969107*	*-1.5383735*
2	30.05667	-30	6.32777884	1	-1.72074942	0.5519951
3	44.84871	-15	2.26928905	1	0.02427581	0.6066289
4	48.39114	-12	3.09862141	1	-0.33231148	0.3575233
5	65.05094	5	1.02886544	1	0.55761964	0.4114555
6	72.18368	12	0.22477223	1	0.90335488	0.8047274
7	73.59252	14	0.85100787	1	0.63409291	0.1669272
8	66.28243	6	3.48076680	1	-0.49662219	-1.1135916
9	73.63474	14	0.41754209	1	0.82046980	0.5912576
10	84.59939	25	0.31853628	1	0.86303924	0.3987047
11	77.10157	17	0.63857190	1	0.72543381	0.2522861
12	75.82544	16	1.85734830	1	0.20139761	-1.0487717
13	83.07405	23	1.91636881	1	0.17602061	-2.2103853
14	*82.83788*	*23*	*0.02513456*	*1*	*0.98919292*	*0.9578935*
15	84.24902	24	0.10587522	1	0.95447693	0.8117228

Figure III.1. Résidus de martingale en fonction de l'âge centré au diagnostic après ajustement d'un modèle exponentiel prenant en compte l'effet de l'âge centré sur le taux de mortalité (M.Agec) et après ajustement d'un modèle exponentiel sans prise en compte de cet effet (M.1)



La figure III.1 représente les résidus de martingale en fonction de l'âge centré. En triangles gris foncé sont représentés les résidus de Martingale (Rm.1) en fonction de l'âge lorsque l'âge n'est pas inclus dans le modèle (M.1) et en ronds sont représentés les résidus de Martingale (Rm.Agec) en fonction de l'âge lorsque l'âge est inclus dans le modèle (M.Agec). Les résidus d'un même individu sont reliés d'un trait vertical en pointillés. Les résidus de martingale ne peuvent pas dépasser 1, d'où la ligne horizontale à ce niveau. Nous pouvons observer que pour les personnes les plus jeunes, Rm.1 < Rm.Agec, alors que pour les personnes les plus vieilles, la tendance est inversée, c'est-à-dire, Rm.1 > Rm.Agec. En effet, cela est dû au fait que pour les jeunes, le taux de mortalité estimé par (M.1) est trop élevé par rapport au taux de mortalité estimé par (M.Agec).

Nous pouvons voir que l'espérance des résidus de Martingale (Rm.1) n'est pas nulle pour l'ensemble des âges (le lissage des résidus a une tendance linéaire), le modèle M.1 n'est donc pas correcte (notons que notre échantillon étant petit, la variabilité est grande, mais ce jeu de données illustre déjà le principe).

Etude de la forme fonctionnelle d'une covariable

Utilisation

Ces résidus permettent particulièrement d'étudier la forme fonctionnelle d'une covariable. En effet, Therneau a démontré [Therneau, 1990] que l'espérance des résidus de martingale était proportionnelle à la forme fonctionnelle des covariables.

Afin d'étudier la forme fonctionnelle d'une covariable, les résidus de martingale peuvent être utilisés graphiquement de la manière suivante :

Les résidus de martingale doivent être représentés en fonction de la covariable d'intérêt [Therneau, 1990] :

- Pour chaque patient i, calculer le résidu de martingale obtenu à partir du modèle ajusté au temps T_i. Nous obtenons ainsi un vecteur de longueur égale au nombre de patients présents dans notre échantillon tel que :

$$\begin{pmatrix} \hat{M}_1(T_1) \\ \hat{M}_2(T_2) \\ \vdots \\ \hat{M}_n(T_n) \\ \end{pmatrix}$$

- Faire le graphique de l'espérance des résidus de martingale (en utilisant un lowess par exemple) en fonction de la covariable d'intérêt z_k.

$$\begin{pmatrix} \hat{M}_1(T_1) \\ \hat{M}_2(T_2) \\ \vdots \\ \hat{M}_n(T_n) \\ \end{pmatrix} \text{ en fonction de } \begin{pmatrix} z_{k1} \\ z_{k2} \\ \vdots \\ z_{kn} \\ \end{pmatrix}$$

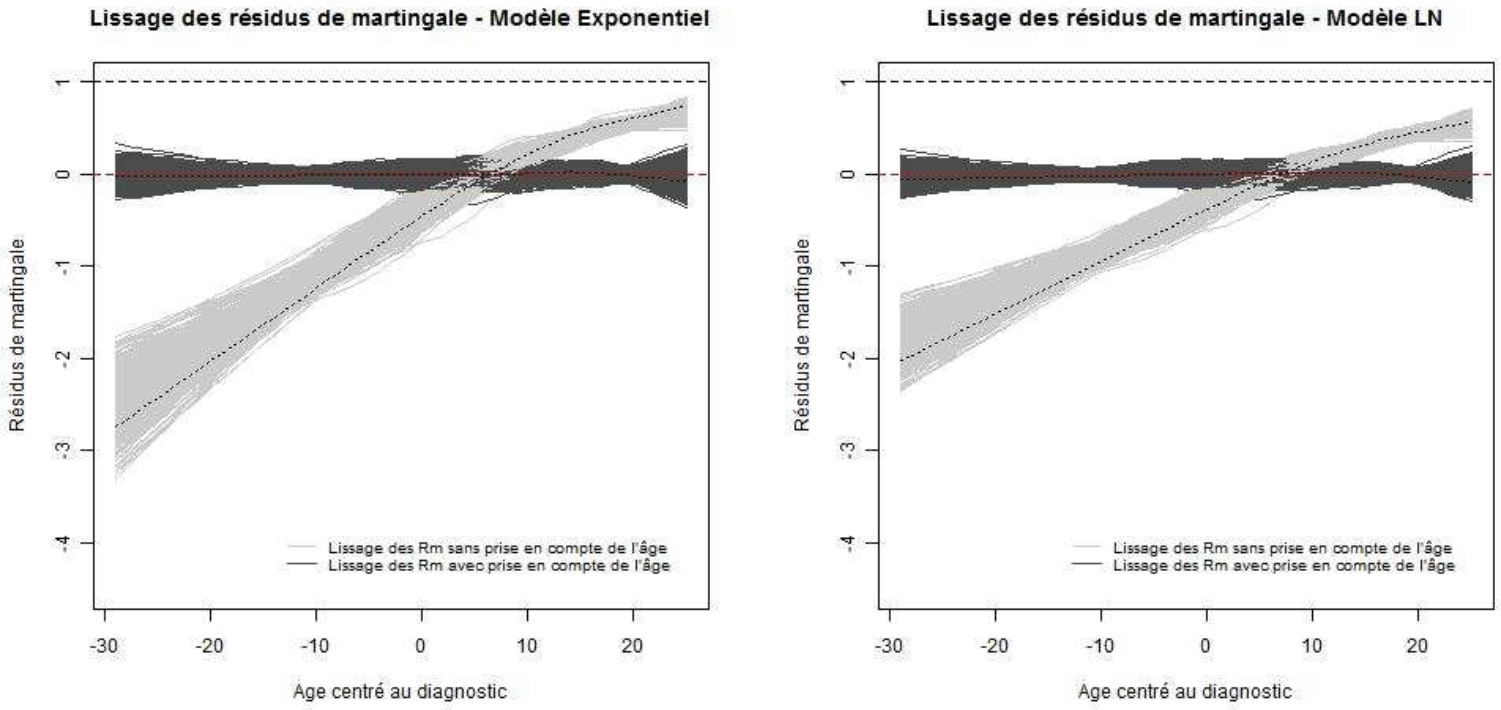
- Si le lissage de ces résidus est une droite horizontale proche de zéro, la covariable d'intérêt a bien été prise en compte dans le modèle
- Si le lissage de ces résidus n'est pas une droite horizontale proche de zéro, cela signifie que la covariable d'intérêt n'a pas bien été prise en compte dans le modèle. Le graphe représente alors une approximation de la forme fonctionnelle de cette covariable.

Exemple

Afin de comprendre leur comportement, nous allons observer la sensibilité des résidus à la forme fonctionnelle modélisée ainsi qu'à la distribution utilisée. Pour cela, nous allons confronter les résidus obtenus après ajustement de différents modèles (modèle exponentiel sans prise en compte de l'effet de l'âge, modèle exponentiel avec prise en compte de l'effet de l'âge, modèle log-normal sans prise en compte de l'effet de l'âge, modèle log-normal avec prise en compte de l'effet de l'âge) sur données simulées. En ce qui concerne ces données, la covariable *âge* suit une loi uniforme qui génère 25% de patients compris dans la classe d'âge [30,65], 35% de patients compris dans la classe d'âge [65,75] et 40% de patients compris dans la classe d'âge [75,85] et la covariable *année de diagnostic* suit une loi uniforme entre 1980 et 1985. La covariable âge aura un effet linéaire-proportionnel sur le taux de mortalité qui s'exprimera alors de la manière suivante : $\lambda(t) = \lambda_{LN}(t) \times \exp(\beta_{Age} \times Agec)$ avec $\beta_{Age} = 0.05$. La covariable *année de diagnostic* suit une loi uniforme entre 1980 et 1985. Le temps de survie des patients T a été généré à l'aide de la méthode de la transformation inverse à partir d'une distribution log-normale avec les paramètres μ et σ respectivement fixés à 0.875 et 1.37. La censure administrative, le temps potentiel de suivi T_{max}, le temps final d'observation et la variable état ont été générés de la même manière qu'exposé dans la partie « Principe ». 1000 échantillons constitués de 500 patients chacun ont été générés.

Les résultats sont les suivants :

Figure III.2. Evaluation de la sensibilité des résidus de martingale à la forme fonctionnelle de l'âge modélisée ainsi qu'à la distribution utilisée



En ce qui concerne le graphe de gauche, le modèle simulé est le modèle log-normal avec prise en compte de la covariable âge de façon linéaire-proportionnelle et les modèles ajustés sont le modèle exponentiel sans prise en compte de la covariable âge dans le modèle (courbes grises) et le modèle exponentiel avec prise en compte de la covariable âge de façon linéaire-proportionnelle (courbes noires).

En ce qui concerne le graphe de droite, le modèle simulé est le modèle log-normal avec prise en compte de la covariable âge de façon linéaire-proportionnelle et les modèles ajustés sont le modèle log-normal sans prise en compte de la covariable âge dans le modèle (courbes grises) et le modèle log-normal avec prise en compte de la covariable âge de façon linéaire-proportionnelle (courbes noires).

Nous pouvons observer que pour les deux distributions étudiées, lorsqu'il n'y a pas d'ajustement sur l'âge, le graphe de la moyenne du lowess des résidus de martingale sur les 1000 jeux de données est linéaire. Cela signifie que la covariable n'a pas bien été prise en compte dans le modèle et qu'elle devrait être prise en compte de manière linéaire. En revanche, lorsqu'il y a eu ajustement sur l'âge, le graphe de la moyenne du lowess des résidus de martingale sur les 1000 jeux de données tend vers une

droite constante proche de zéro. Cela signifie que la covariable a bien été prise en compte dans le modèle.
Les résidus de martingale prennent donc bien en compte la façon dont a été intégrée la covariable d'intérêt dans le modèle mais ne sont pas sensibles à la distribution utilisée car les conclusions que l'on peut faire sont similaires. En effet, lorque que l'on ajuste un modèle exponentiel avec l'effet de l'âge correctement spécifié, les résidus ne détectent pas que la distribution utilisée n'est pas correcte (les données sont simulées avec une loi log-normale et non pas une distribution exponentielle).

III.3.2 Résidus de Schoenfeld

Définition

Les résidus de Schoenfeld [Schoenfeld, 1982] ont été définis uniquement dans le cadre du modèle de Cox, pour chaque covariable incluse dans le modèle, à partir de la vraisemblance partielle.

En ne raisonnant que sur les r patients décédés, la probabilité conditionnelle que cela soit le patient j qui décède au temps t_j, sachant les personnes à risque en t_j, est (vraisemblance partielle individuelle) :

$$\upsilon_j(\beta) = P\left(D_j \middle| R_j\right) = \frac{\lambda_0(t_j)\exp(\beta z_j)\times dt}{\sum_{l\in R_j}\lambda_0(t_j)\exp(\beta z_l)\times dt} = \frac{\exp(\beta z_j)}{\sum_{l\in R_j}\exp(\beta z_l)}$$

où D_j représente l'évènement décès pour le patient j et R_j représente le nombre de patients à risque au temps t_j.

La vraisemblance partielle globale s'exprime alors de la manière suivante (avec son équivalent dans le cadre des processus de comptage) :

$$V^*(\beta) = \prod_{i=1}^{r} \upsilon_i(\beta) = \prod_{i=1}^{r} \frac{\exp(\beta z_i)}{\sum_{j \in R_i} \exp(\beta z_j)} \qquad \Leftrightarrow \qquad V^*(\beta) = \prod_{i=1}^{n} \prod_{t \geq 0} \left[\frac{Y_i(t)\exp(\beta z_i)}{\sum_j Y_j(t)\exp(\beta z_j)} \right]^{dN_i(t)}$$

En passant à la log-vraisemblance (τ temps maximal de suivi),

$$\log V^*(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \log \left[\left(\frac{Y_i(u)\exp(\beta z_i)}{\sum_j Y_j(u)\exp(\beta z_j)} \right)^{dN_i(u)} \right]$$

$$\log V^*(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left[\log(Y_i(u)\exp(\beta z_i)) - \log\left(\sum_j Y_j(u)\exp(\beta z_j) \right) \right] dN_i(u)$$

$$\log V^*(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left[\beta z_i - \log\left(\sum_j Y_j(u)\exp(\beta z_j) \right) \right] dN_i(u)$$

Afin d'estimer les paramètres du modèle, le maximum de vraisemblance est utilisé. La $k^{\text{ième}}$ composante du vecteur score, qui représente la dérivée de la log-vraisemblance au temps maximal de suivi τ par rapport à la variable β_k, est la suivante (avec $k \in \{1,...,p\}$, p étant le nombre de covariables incluses dans le modèle) :

$$U_k(\beta,\tau) = \frac{d\log\left(V^*(\beta)\right)}{d\beta_k}$$

$$U_k(\beta,\tau) = \sum_i \int_0^{\tau} \left[z_{ik} - \frac{\sum_j Y_j(u) z_{jk} \exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)} \right] dN_i(u)$$

L'utilisation du score partiel nous amène à exprimer la moyenne pondérée de la $k^{\text{ième}}$ covariable au cours du temps par :

$$\bar{z}_k(\beta,t)=\frac{\sum_j Y_j(t)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(t)\exp(\beta z_j)}$$

et la variance pondérée de la $k^{\text{ième}}$ covariable au cours du temps par :

$$V(\beta,t)=\frac{\sum_j Y_j(t)z_{jk}^2\exp(\beta Z_j)}{\sum_j Y_j(t)\exp(\beta z_j)}-\left[\frac{\sum_j Y_j(t)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(t)\exp(\beta z_j)}\right]^2$$

Les résidus de Schoenfeld du patient décédant au temps t_i, Sch_i, sont définis comme un vecteur de p composantes $Sch_i = (Sch_{i1},\ldots, Sch_{ip})$ tel que :

$$Sch_{ik}=\int_0^{t_i}[z_{ik}-\bar{z}_k(\beta,u)]dN_i(u) \qquad \Leftrightarrow \qquad Sch_{ik}=[z_{ik}-\bar{z}_k(\beta,t_i)] \qquad \text{(III.4)}$$

Ces résidus représentent la différence entre la valeur observée de la $k^{\text{ième}}$ covariable de l'individu décédant au $i^{\text{ième}}$ temps et la valeur attendue de cette covariable pour le patient décédé sous le modèle proportionnel ajusté. Cette valeur attendue est une moyenne pondérée de la covariable z_k par le risque de décès des patients encore à risque. Si l'effet de la covariable d'intérêt est proportionnel, la différence ne doit pas dépendre du temps et doit fluctuer autour de zéro. Si la différence est importante, cela signifie que l'individu s'éloigne du profil moyen ; dans certains cas, cela est dû au fait que le patient décède trop tard ou trop tôt par rapport à son risque de décès.

Test de l'hypothèse des taux proportionnels

Lien entre les résidus de Schoenfeld et le coefficient de la covariable d'intérêt

Ces résidus ont été développés afin de pouvoir tester l'hypothèse des taux proportionnels du modèle de Cox ; c'est-à-dire, étudier si l'effet est constant ou s'il varie au cours du temps [Schoenfeld, 1982][Grambsh, 1994].

Les résidus de Schoenfeld sont définis à chaque temps d'évènement de la manière suivante :

$$Sch_{ik} = \{z_{ik} - \bar{z}(\beta, t_i)\}$$
$$Sch_{ik} = \{z_{ik} - \bar{z}(\beta(t_i), t_i)\} + \{\bar{z}(\beta(t_i), t_i) - \bar{z}(\beta, t_i)\}$$

En utilisant le développement de Taylor, le second terme s'exprime de la manière suivante :

$$\bar{z}(\beta(t_i), t_i) - \bar{z}(\beta, t_i) \approx (\beta(t_i) - \beta)\frac{\partial}{\partial \beta}(\bar{z}(\beta, t_i))$$

Le premier terme représentant les résidus de Schoenfeld sous le vrai modèle, son espérance est nulle. L'espérance de ces résidus conditionnellement à ce qui a été observé jusqu'en t_i (dénoté par F_{ti}) s'exprime alors de la manière suivante :

$$E\left(Sch_{ik} \middle| F_{t_i}\right) \approx E\left(\{z_{ik} - \bar{z}(\beta(t_i), t_i)\} \middle| F_{t_i}\right) + (\beta(t_i) - \beta)\frac{\partial}{\partial \beta}(\bar{z}(\beta, t_i))$$
$$E\left(Sch_{ik} \middle| F_{t_i}\right) \approx 0 + (\beta(t_i) - \beta)\frac{\partial}{\partial \beta}(\bar{z}(\beta, t_i))$$

La forme de l'effet de la covariable d'intérêt au cours du temps peut être approximée par :

$$\beta(t_i) \approx \beta + E\left(Sch_{ik} \middle| F_{t_i}\right)\left[\frac{\partial}{\partial \beta}(\bar{z}(\beta, t_i))\right]^{-1}$$

Or,

$$\frac{\partial}{\partial\beta}\left(\bar{z}(\beta,t_i)\right)=\frac{\sum_j Y_j(t_i)z_{jk}^2\exp(\beta z_j)}{\sum_j Y_j(t_i)\exp(\beta z_j)}-\left[\frac{\sum_j Y_j(t_i)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(t_i)\exp(\beta z_j)}\right]^2$$

$$\frac{\partial}{\partial\beta}\left(\bar{z}(\beta,t_i)\right)=V(\beta,t_i)$$

Donc, comme cela a été démontré [Grambsch, 1994],

$$E\left(Sch_{ik}^{*}\middle|F_{t_i}\right)+\beta\approx\beta(t_i)$$

Sch_{ik}^{*} représentant le résidu de Schoenfeld standardisé de la $k^{\text{ième}}$ covariable pour le patient décédant au temps t_i.

L'approximation faite par Schoenfeld [Schoenfeld, 1982] est équivalente :

$$E(Sch_{ik})\approx g_k(t_i)\times\left(E\left(z_{ik}^2\middle|R_i\right)-E\left(z_{ik}\middle|R_i\right)^2\right)$$

avec z_{ik} représentant la $k^{\text{ième}}$ covariable de l'individu décédant au temps t_i et $g_k(t)$ étant fonction de β_k et $\beta_k(t)$

Ces deux approximations montrent que toutes les variations de l'effet de la covariable d'intérêt en fonction du temps peuvent être identifiées en faisant le graphe des résidus de Schoenfeld en fonction du temps. Leur somme sur tous les patients étant équivalente au vecteur score, elle doit être égale à zéro :

$$U_k(\beta)=\sum_{i=1}^{n}Sch_{ik}=0$$

Evolution de la moyenne pondérée

L'objectif dans cette partie est d'étudier l'évolution de la moyenne pondérée dans le cas simple d'une variable binaire, ce qui correspond à étudier l'évolution du second terme dans l'expression des résidus de Schoenfeld :

$$Sch_{ik} = \left\{ z_{ik} - \frac{\sum_{j \in R_i} z_{jk} \exp(\beta z_j)}{\sum_{j \in R_i} \exp(\beta z_j)} \right\}$$

En effet, la manière dont va évoluer la moyenne pondérée va permettre la détection du caractère proportionnel ou non de l'effet de la covariable z_k.

Cette étude va s'effectuer à l'aide de deux scénarios :

◦ *Scénario 1*

L'effet de la covariable (ici, le *sexe*) est proportionnel (PH).

50% Femmes : $\lambda_F(t) = \lambda_0(t)\exp(\beta_{Sexe} \times Sexe) = 0.5 \times \exp(-0.25 \times 1) = 0.3894$
50% Hommes : $\lambda_H(t) = \lambda_0(t)\exp(\beta_{Sexe} \times Sexe) = 0.5 \times \exp(-0.25 \times 0) = 0.5$

◦ *Scénario 2*

L'effet du sexe est non-proportionnel (NPH) : le taux relatif homme *vs* femme est 0.5/0.3894 entre 0 et 5 ans, puis 0.25/0.3894 entre 5 et 15 ans.

50% Femmes : $\lambda_F(t) = \lambda_0(t)\exp(\beta_{Sexe} \times Sexe) = 0.5 \times \exp(-0.25 \times 1) = 0.3894$
50% Hommes : $t \leq 5$: $\lambda_H(t) = \lambda_0(t)\exp(\beta_{Sexe} \times Sexe) = 0.5 \times \exp(-0.25 \times 0) = 0.5$
$t > 5$: $\lambda_H(t) = 0.5 \times \lambda_0(t)\exp(\beta_{Sexe} \times Sexe) = 0.25 \times \exp(-0.25 \times 0) = 0.25$

◦ *Proportions théoriques et survies théoriques*

Soient $S_F(t)$ (femmes) et $S_H(t)$ (hommes) les survies de chaque groupe au temps *t*.
Soient $\alpha_F(t)$ (femmes) et $\alpha_H(t)$ (hommes) les proportions de patients dans chaque groupe au temps *t*.

Les taux de mortalités étant générés à l'aide d'une distribution exponentielle (constants au cours du temps), la survie de chaque groupe s'exprime de la manière suivante :

$$S_F(t) = \exp\left(-\int_0^t \lambda_F du\right) = \exp(-\lambda_F t)$$

$$S_H(t) = \begin{cases} \exp\left(-\int_0^t \lambda_{H[0;5]} du\right) = \exp(-\lambda_{H[0;5]} \times t) & \text{Si } t \le 5 \\ \exp\left(-\int_0^5 \lambda_{H[0;5]} du - \int_5^t \lambda_{H[5;15]} du\right) = \exp(-\lambda_{H[0;5]} \times 5 - \lambda_{H[5;15]} \times (t-5)) & \text{Si } t > 5 \end{cases}$$

En ce qui concerne les proportions de femme ou d'homme encore à risque au temps *t* :

En t = 0,

$$\alpha_F(0) = 0.5$$
$$\alpha_H(0) = 0.5$$

$\forall t \in R$,

$$\alpha_F(t) = \frac{\alpha_F(0) S_F(t)}{\alpha_F(0) S_F(t) + \alpha_H(0) S_H(t)}$$
$$\alpha_H(t) = \frac{\alpha_H(0) S_H(t)}{\alpha_F(0) S_F(t) + \alpha_H(0) S_H(t)}$$

◦ *Moyennes pondérées théoriques*

A partir des proportions théoriques, on peut donc calculer les moyennes pondérées théoriques à chaque temps :

En partant de la formule générale, on a :

$$\bar{z}(\beta_{sexe}, t) = \frac{\sum_{j \in R_i} sexe_j \exp(\beta_{sexe} sexe_j)}{\sum_{j \in R_i} \exp(\beta_{sexe} sexe_j)}$$

$$\bar{z}(\beta_{sexe}, t) = \frac{n_H(t) \times (0 \times \exp(\beta_{sexe} \times 0)) + n_F(t) \times (1 \times \exp(\beta_{sexe} \times 1))}{n_H(t) \times (\exp(\beta_{sexe} \times 0)) + n_F(t) \times (\exp(\beta_{sexe} \times 1))}$$

Puis en passant par les proportions, la formule devient :

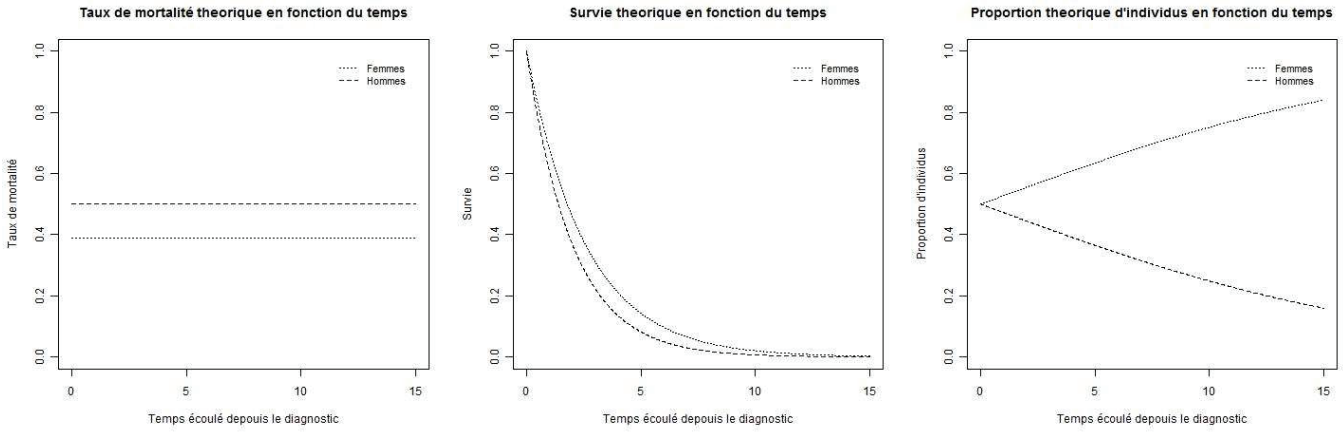
$$\bar{z}(\beta_{sexe}, t) = \frac{\alpha_H(t) \times (0 \times \exp(\beta_{sexe} \times 0)) + \alpha_F(t) \times (1 \times \exp(\beta_{sexe} \times 1))}{\alpha_H(t) \times (\exp(\beta_{sexe} \times 0)) + \alpha_F(t) \times (\exp(\beta_{sexe} \times 1))}$$

◦ *Graphes des quantités théoriques (figure III.3-III.4)*

En observant les taux de mortalité théoriques du premier scénario, les hommes ayant un taux de mortalité plus importants que les femmes (figure III.3.a), leur survie diminue plus rapidement (figure III.3.b). La proportion d'homme au cours du temps doit alors diminuer alors que celle des femmes doit augmenter (figure III.3.c). Nous nous attendons alors à ce que la moyenne pondérée tende vers la valeur 1 (vers les femmes) (figure III.4.a). En ce qui concerne le deuxième scénario (figure III.3.d), dans un premier temps, de 0 à 5 ans, le scénario est identique au précédent. Dans un second temps, à partir de 5 ans de suivi, le taux de mortalité des hommes diminue et devient inférieur à celui des femmes (la cassure est due au fait que l'effet théorique du sexe est une fonction par morceaux). Leur survie re-augmente (figure III.3.e), tout comme leur proportion au sein du groupe (figure III.3.f). La moyenne pondérée tend donc vers celle des hommes, c'est-à-dire zéro (figure III.4.b).

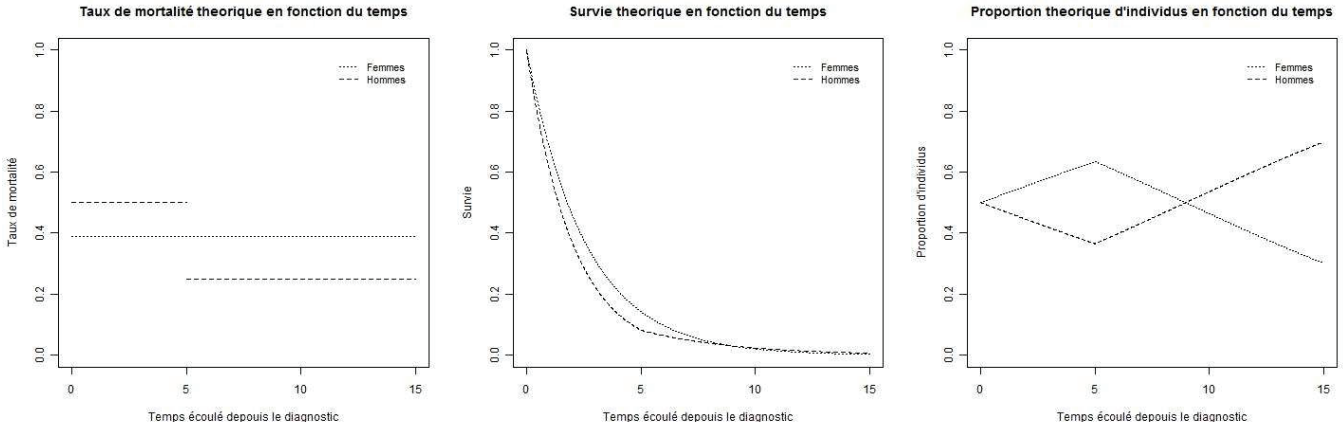
Figure III.3 : Grandeurs théoriques sous les différents scénarios analysés pour l'étude du caractère proportionnel de la covariable *sexe*

Scénario 1 : Données simulées LIN-PH



(a) (b) (c)

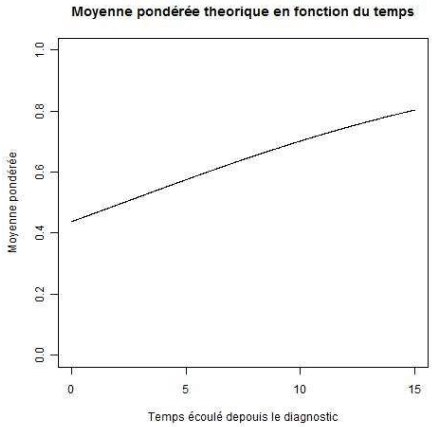
Scénario 2 : Données simulées LIN-NPH



(d) (e) (f)

Figure III.4 : Moyennes pondérées théoriques sous les différents scénarios analysés pour l'étude du caractère proportionnel de la covariable sexe

(a) Scénario 1 : LIN-PH



(b) Scénario 2 : LIN-NPH

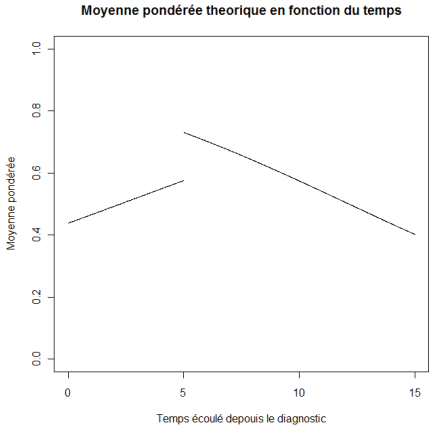
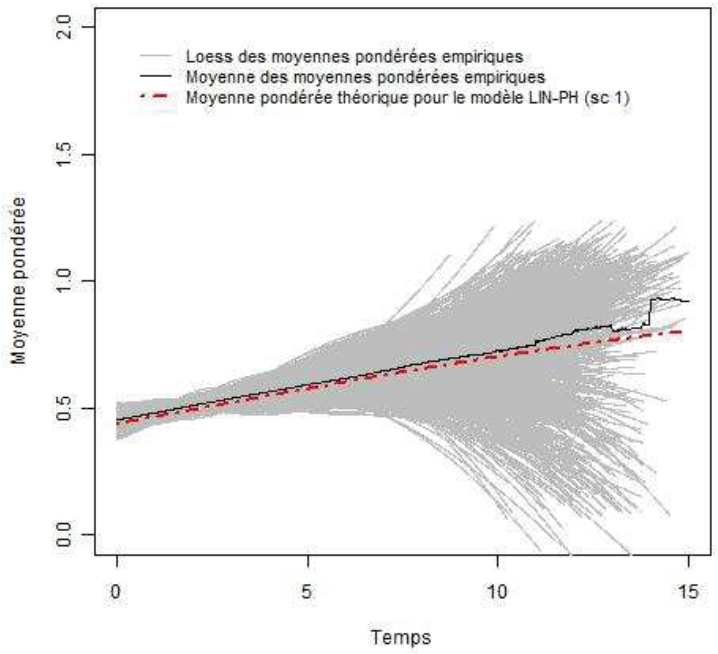
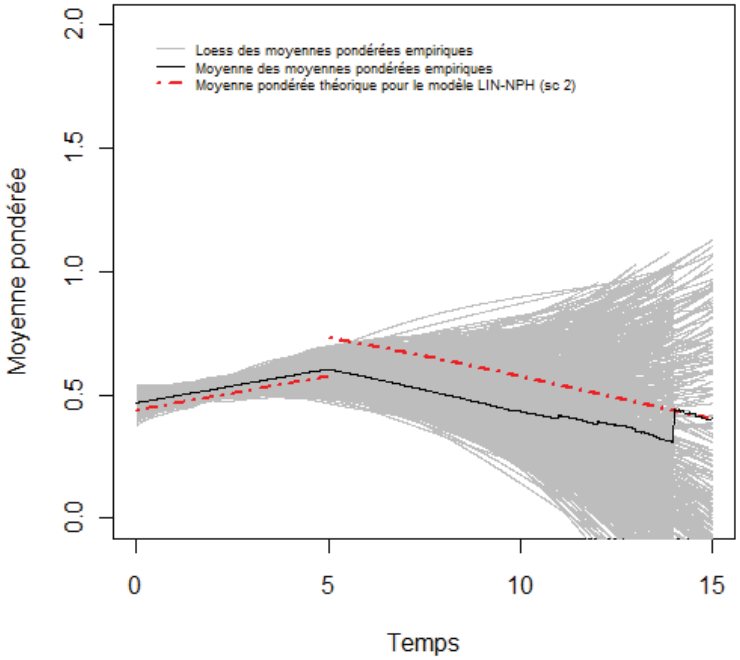


Figure III.5 : Moyennes pondérées empiriques sous les différents scénarios analysés pour l'étude du caractère proportionnel de la covariable sexe

Scénario 1



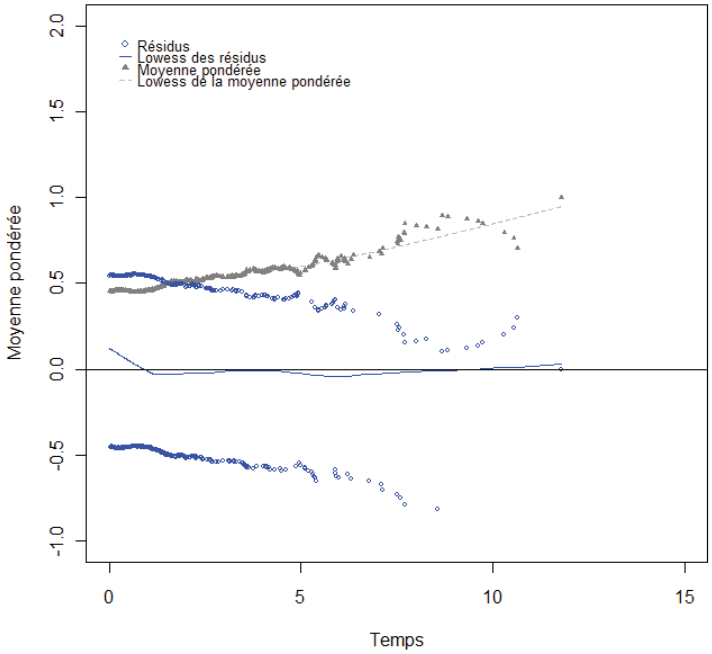
Scénario 2



En ajustant un modèle proportionnel sur les données obtenues à l'aide des deux scénarios étudiés, nous allons observer l'allure de la moyenne des moyennes pondérées empiriques (Figure III.5). Lorsque nous nous plaçons dans le scénario 1, nous pouvons observer que la moyenne des moyennes pondérées empiriques sur les 1000 échantillons suit bien l'allure de la moyenne pondérée théorique. La mauvaise adéquation sur la fin du suivi (à partir de 11 ans environ) est due au fait que pour beaucoup de jeu de données, le dernier décès a eu lieu avant, la moyenne pondérée n'est donc plus estimée au-delà (en moyenne, il reste 3.047 personnes à risque à onze ans de suivi). La figure III.6.a, qui est une illustration des résidus de Schoenfeld sur un jeu de données simulé dans le cadre du scénario 1, montre bien que le lissage des résidus est proche de zéro et ne reflète aucune forme particulière de l'effet de l'âge en fonction du temps. En revanche, pour le second scénario, la moyenne des moyennes pondérées empiriques ne suit pas la moyenne pondérée théorique du modèle LIN-PH mais bien la moyenne pondérée théorique du modèle LIN-NPH. Sur la première période, de 0 à 5 ans, nous nous attendons à ce que la moyenne pondérée empirique soit confondue avec la moyenne pondérée théorique. L'écart que nous pouvons observer est dû au fait que lorsque l'on ajuste le modèle proportionnel sur les données, l'effet de du sexe estimé est biaisé, ce qui conduit à une moyenne pondérée biaisée. En ce qui concerne la seconde période, à partir de 5 ans de suivi, la moyenne pondérée empirique tend vers zéro comme attendue (la mauvaise adéquation à la fin du suivi ainsi que l'écart à la moyenne pondérée théorique sont dus aux mêmes raisons que celles expliquées précédemment, à onze ans de suivi, il reste en moyenne 5.457 personnes à risque). La figure III.6.b illustre l'utilisation des résidus de Schoenfeld sur un jeu de données simulé dans le cadre du scénario 2. Contrairement à la figure III.6.a, nous pouvons observer que le lissage des résidus n'est pas proche de zéro : l'effet du sexe varie au cours du temps.

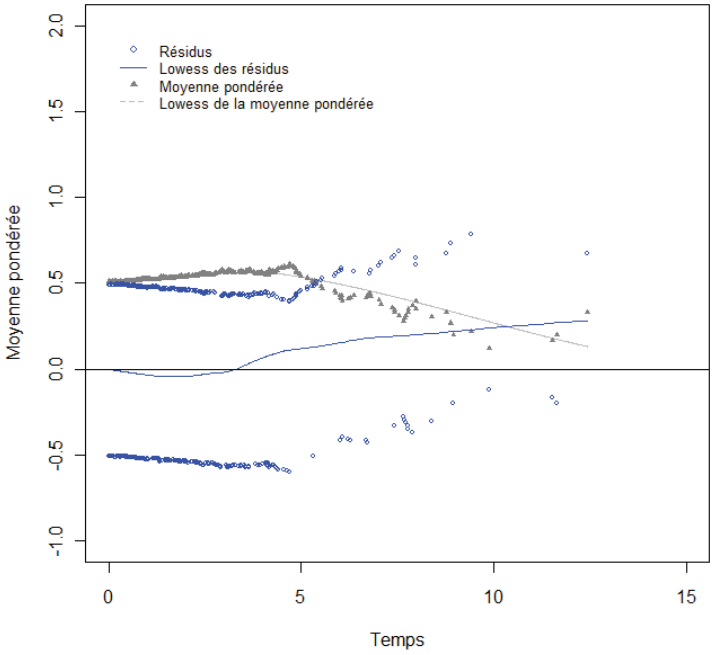
Figure III.6 : Illustration des résidus de Schoenfeld et de la moyenne pondérée associée sous les différents scénarios analysés pour l'étude du caractère proportionnel de la covariable sexe

Scénario 1



(a)

Scénario 2



(b)

Dans le cas d'une variable binaire, nous avons pu voir que lorsque la moyenne pondérée ne change pas d'allure au cours du temps, cela signifie que l'effet de la covariable est proportionnel, sinon l'effet est non-proportionnel, la répartition de la covariable d'intérêt au cours du temps n'évolue pas comme attendu dans le cadre d'un modèle proportionnel.

III.3.3 Transformées de Martingale

La théorie des processus stochastiques a rendu disponibles de nombreux résultats utiles à l'élaboration d'outils permettant de tester les hypothèses des modèles de survie. Les résultats les plus utiles dans notre cas sont liés aux transformées de martingale *MT* (Martingale Transformed), qui s'expriment de façon générale comme suit :

$$MT(t) = \sum_{i=1}^{n} \int_0^t h_i(u)\,dM_i(u) \qquad \text{(III.5)}$$

avec *i* étant l'indice du patient, h_i étant un processus dit « prévisible » associé au patient *i* et M_i la martingale associée au patient *i*.

Les transformées de martingale sont elles-mêmes des martingales et conservent donc les propriétés des martingales. On sait donc qu'une transformée de martingale a une espérance nulle et une variance égale à E($< M >_t$).

Le fait d'exprimer certaines quantités en tant que transformées de martingale permet de connaitre leur distribution sous l'hypothèse nulle et donc de voir quand est-ce que celle-ci peut être acceptée ou rejetée.

Dans le cadre des modèles de survie, en posant $h_i(t)$ un processus prévisible, on peut définir l'intégrale de Stieljes aléatoire du processus h_i par rapport au processus M_i telle que :

$$\int_0^t h_i(u)\,d\hat{M}_i(u) = \int_0^t h_i(u)\,dN_i(t) - \int_0^t h_i(u)\,d\hat{\Lambda}_i(t)$$

Dans cette expression, nous pouvons retrouver le terme observé $\int_0^t h_i(u)\,dN_i(t)$, le terme prédit par le modèle $\int_0^t h_i(u)\,d\hat{\Lambda}_i(t)$ ainsi que le terme résiduel $\int_0^t h_i(u)\,d\hat{M}_i(u)$. Ce dernier terme mesure une différence entre ce qui est observé et ce qui est prédit par le modèle, nous pouvons alors bien parler de résidus.

La plupart des résidus sont utilisés à travers des tests graphiques, sans test formel. Les résultats obtenus doivent donc être interprétés avec précaution. En effet, il est souvent difficile de déterminer si les tendances qui en ressortent sont dues à la spécification de mauvaises hypothèses de départ lors de la construction de modèle ou si cela est dû à de simples variations aléatoires. Il est donc plus prudent d'utiliser des résidus dont les caractéristiques sous l'hypothèse nulle (H_0 : l'hypothèse testée est correcte), sachant les autres hypothèses correctes, sont connues.

Les résidus, présentés dans la suite, ont été proposés par Lin [Lin, 1993][Lin, 1996] sous la forme générale des processus suivants :

Soit z_k la covariable d'intérêt.

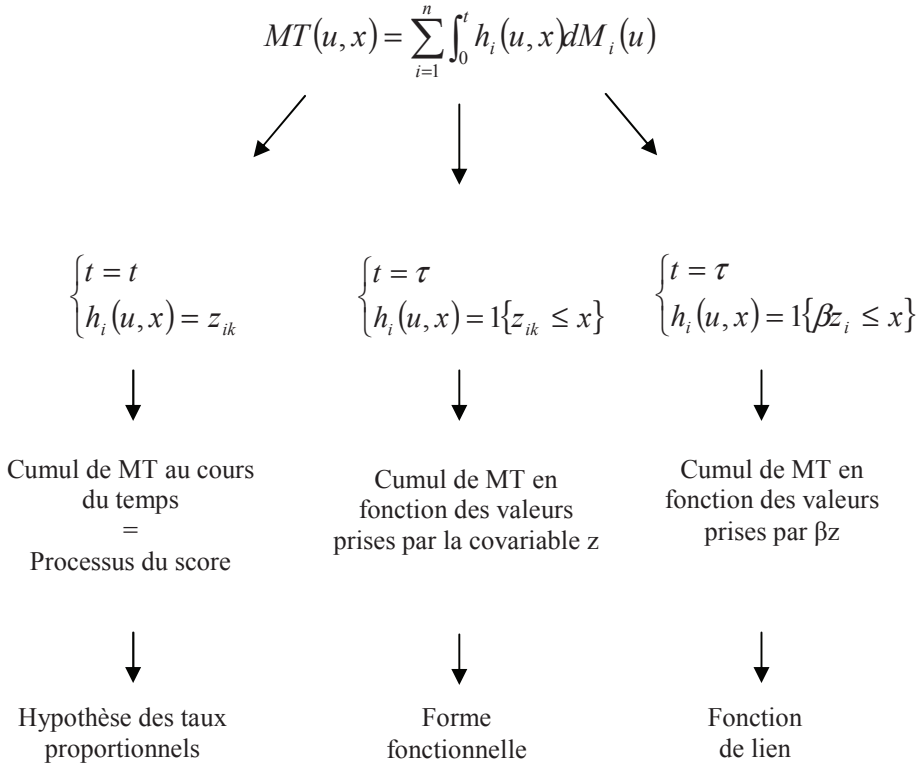
$$W_z(t, z_k) = n^{-1/2} \sum_{i=1}^{n} f(z_{ik})\, I(z_{ik} \le x) M_i(t)$$

$$W_r(t, z_k) = n^{-1/2} \sum_{i=1}^{n} f(z_{ik})\, I(\beta z_i \le x) M_i(t)$$

Le processus du score qui permet de tester l'hypothèse de proportionnalité de l'effet de la covariable d'intérêt est un cas particulier du processus W_z en prenant f(y) = y et x = $+\infty$.
Les résidus de martingale cumulés qui permettent de vérifier si la forme fonctionnelle est correcte sont un cas particulier du processus W_z en prenant f(y) = 1 et $t = \tau$.
Les résidus de martingale cumulés qui permettent de vérifier si la fonction de lien est correcte sont un cas particulier du processus W_r en prenant f(y) = 1 et $t = \tau$.

Nous avons pu voir que ces processus pouvaient s'écrire sous la forme d'une transformée de martingale: en utilisant différentes fonctions h_i, différentes hypothèses peuvent être testées comme présenté ci-dessous (figure III.7) :

Figure III.7. Trois choix de fonction h_i pour définir trois transformées de martingale différentes pour tester les trois principales hypohèses en analyse de survie



III.3.3.a. Processus du score

Dans cette sous-partie, nous allons voir que le processus du score peut s'écrire comme une transformée de martingale. En utilisant cette écriture, un test formel reposant sur le processus du score et utilisant les propriétés des martingales pourra donc être élaboré pour tester l'hypothèse des taux proportionnels.

Définition

En se plaçant sous le modèle (III.1) que nous rappelons ici :

$$\lambda_i(t) = \lambda_0(t)\exp(\beta z_i)$$

le processus du score au temps t représente le vecteur de la dérivée de la log-vraisemblance au temps t par rapport à β_k, $t \in [0,\tau]$ (τ représentant le temps de suivi maximal dans la population étudiée), β_k coefficient des covariables du modèle, $k \in \{1,\ldots,p\}$, p nombre de covariables dans le modèle.

$$U(\beta,t) = \begin{pmatrix} \dfrac{\partial \log V(\beta,t)}{\partial \beta_1} \\ \vdots \\ \dfrac{\partial \log V(\beta,t)}{\partial \beta_k} \end{pmatrix}$$

Le processus du score ne s'exprime pas de la même manière dans le modèle de Cox et dans un modèle paramétrique.

Modèle de Cox

Dans le cadre du modèle de Cox [Cox, 1972], le processus du score pour la $k^{\text{ième}}$ covariable peut s'exprimer de la façon suivante (Appendix 1) :

$$U_k(\beta,t)=\sum_i\int_0^t\left[z_{ik}-\frac{\sum_j Y_j(u)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)}\right]dM_i(u)$$

Le processus du score est donc un cas particulier des résidus de transformée de martingale où la fonction prévisible h_i est égale à $[z_{ik}-\bar{z}(\beta,u)]$.

Modèle paramétrique

Dans le cadre d'un modèle paramétrique, la log-vraisemblance s'écrit de la manière suivante (Appendix 2.1) :

$$L(\beta,t)=\sum_i[-\Lambda_i(t)+\delta_i\log(\lambda_i(t))]$$

La dérivée de la log-vraisemblance par rapport à β_k au temps t, qui représente le processus du score pour la $k^{\text{ième}}$ covariable au temps t, s'écrit de la manière suivante (Appendix 2.1) :

$$U_k(\beta,t)=\sum_i\left[\int_0^t z_{ik}dM_i(u)\right]$$

Le processus du score est donc un cas particulier des résidus de transformée de martingale où la fonction prévisible h_i est égale à z_{ik}.

Le fait d'écrire le processus du score comme une transformée de martingale permet de déduire sa distribution sous l'hypothèse nulle : le processus doit fluctuer autour de zéro avec une certaine variance.

Test de l'hypothèse des taux proportionnels

Intuition

Le score est évalué à chaque temps de suivi $t = (t_1,\ldots,t_m)$, comme si les données étaient artificiellement censurées en $t = (t_1,\ldots,t_m)$. Or si le modèle proportionnel est correct,

l'estimation des coefficients est identique que l'on censure les données à t_1 ou t_2 ou à n'importe quel t. Donc sous un modèle proportionnel, si on évalue le processus du score à n'importe quel temps t, il doit être proche de zéro car cela correspond au score qu'on aurait si on censurait les données en t. Ce processus doit donc fluctuer autour de zéro avec une variance égale à $\mathrm{E}\langle U_k \rangle_t$.

Approximation de la distribution du processus du score sous l'hypothèse nulle

En utilisant le développement de Taylor pour le vecteur β ($\beta = \{\beta_1, ..., \beta_p\}$, p étant le nombre de covariables introduites dans le modèle) (Appendix 2.2), il a été établi [Lin, 1996] que la distribution de $n^{-1/2}U(\hat{\beta},t)$ etait asymptotiquement équivalente à la distribution du processus W_z, tel que :

$$W_z(t) = n^{-1/2}\left(U(\beta_0;t) - I(\hat{\beta};t)I(\hat{\beta};\tau)^{-1}U(\beta_0;\tau)\right)$$

$U(\beta_0,t)$ représentant le processus du score sous l'hypothèse nulle au temps t, $U(\beta_0,\tau)$ représentant le processus du score sous l'hypothèse nulle au temps τ (score usuel), $I(\beta,t)$ représentant l'opposé de la dérivée seconde du logarithme de la vraisemblance au temps t et $I(\beta,\tau)$ représentant l'opposé de la dérivée seconde du logarithme de la vraisemblance au temps τ (matrice information).

Cependant, nous connaissons la distribution du processus du score observé $U(\hat{\beta},t)$ mais pas celle du processus du score $U(\beta_0,t)$ sous l'hypothèse nulle.

En approximant M_i par le processus gaussien $N_i(u).G_i$ (Appendix 2.2), G_i suivant une loi normale standardisée, $U_k(\beta_0;t)$ pourra alors être approximé par $\hat{M}_{1k}(t)$ tel que pour le modèle de Cox, $\hat{M}_{1k}(t)$ est égale à :

$$\hat{M}_{1k}(t) = \sum_i \int_0^t \left[z_{ik} - \frac{\sum_j Y_j(u) z_{jk} \exp\left(\hat{\beta} z_j\right)}{\sum_j Y_j(u) \exp\left(\hat{\beta} z_j\right)} \right] dN_i(u) G_i$$

et pour le modèle paramétrique de type (III.1), $\hat{M}_{1k}(t)$ est égale à :

$$\hat{M}_{1k}(t) = \sum_i \int_0^t z_{ik} dN_i(u) G_i$$

La distribution de $W_z(t,z)$ étant équivalentes asymptotiquement au processus $\tilde{W}_z(t,z)$ (Appendix 2.3), le processus du score peut donc être approximé par :

$$\tilde{W}_z(t,z) = n^{-1/2}\left(\hat{M}_1(t) - I\left(\hat{\beta};t\right) I\left(\hat{\beta};\tau\right)^{-1} \hat{M}_1(\tau) \right) \qquad \text{(III.6)}$$

Test statistique pour tester l'hypothèse H_0 : « L'hypothèse des taux proportionnels est correcte »

Pour tester l'hypothèse H_0 : « L'hypothèse des taux proportionnels est correcte » (en supposant les autres hypothèses du modèle correctes), la statistique de test utilisée est la valeur maximale (en valeur absolue) du processus du score observé :

$$\sup_t \left| n^{1/2} U_k\left(\hat{\beta},t\right) \right|$$

Comme classiquement en statistique, si cette statistique de test est « trop élevée » par rapport à ce qui est attendu sous H_0, alors nous aurons tendance à rejeter H_0.

D'un point de vue pratique, des processus gaussiens (III.6) dont la distribution sous l'hypothèse nulle est équivalente à celle du processus $n^{1/2} U_k\left(\hat{\beta},t\right)$ seront simulés.

Afin d'accepter ou de rejeter l'hypothèse nulle, la p-value, représentant le nombre de fois pour lesquelles la statistique de test est inférieure à la valeur maximale (en valeur absolue) des processus gaussiens simulées sera estimée :

$$p.val = \frac{\sum_i \left(\sup_x \left| n^{1/2} U_k\left(\hat{\beta}, t\right) \right| < \sup \left| sim_i \right| \right)}{n.sim}$$

Avec *n.sim* représentant le nombre de processus gaussiens simulés et $\left|sim_i\right|$ représentant le processus gaussien *i* en valeur absolue.

Autrement dit, en fixant α à 5%, l'hypothèse de proportionnalité sera rejetée si la statistique de test est supérieure au maximum des processus simulés dans plus de 95% des cas.

Il peut être utile de produire le graphe associé à ce test formel afin d'observer à quel moment du suivi nous nous éloignons de l'hypothèse proportionnel.
Il faut noter que ce graphe ne permet pas de tester formellement l'hypothèse H_0 (à part dans des cas particulier tel que le maximum du processus observé est supérieur au maximum des maximum des processus gaussien simulés).

Le processus du score pour la covariable *k* doit être représenté en fonction du temps :

- Pour chaque patient *i*, calculer la contribution au score individuel pour la covariable *k* à chaque temps de suivi distinct obtenu à partir du modèle ajusté. Nous obtenons ainsi une matrice de dimension (*n.m*) avec *n* étant le nombre de patients et *m* étant le nombre de temps de suivi distincts.

$$\begin{pmatrix} U_{1k}\left(\hat{\beta}, t_1\right) & U_{1k}\left(\hat{\beta}, t_2\right) & \cdots & \cdots & U_{1k}\left(\hat{\beta}, t_m\right) \\ \vdots & \cdots & & & \vdots \\ \vdots & & \cdots & & \vdots \\ \vdots & & & \cdots & \vdots \\ U_{nk}\left(\hat{\beta}, t_1\right) & U_{nk}\left(\hat{\beta}, t_2\right) & \cdots & \cdots & U_{nk}\left(\hat{\beta}, t_m\right) \end{pmatrix}$$

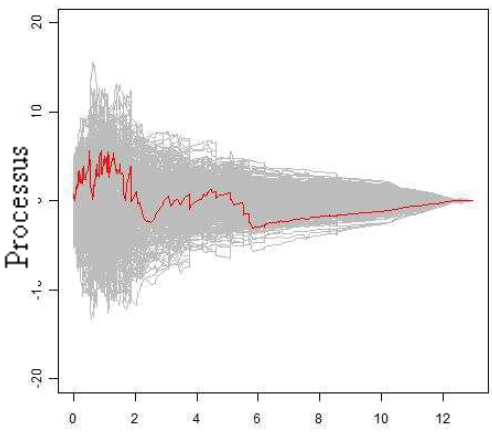
- Sommer les scores individuels pour la covariable *k* à chaque temps pour avoir la valeur du processus du score pour la covariable *k* à chaque temps et obtenir un vecteur de taille *m* tel que :

$$\left(U_k\left(\hat{\beta},t_1\right),U_k\left(\hat{\beta},t_2\right),\cdots,U_k\left(\hat{\beta},t_m\right)\right)$$

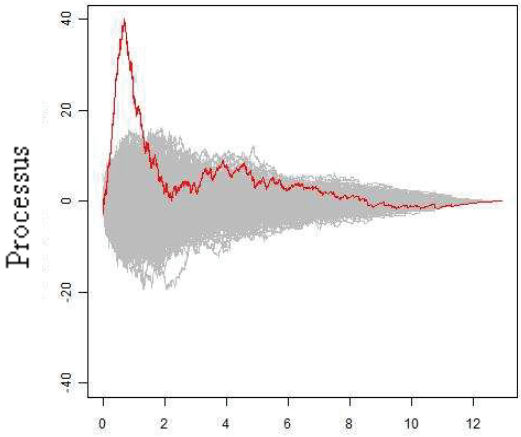
- Faire le graphique du processus du score pour la covariable *k* en fonction des m temps.
- Faire le graphique de plusieurs processus gaussiens correspondants en fonction du temps.

Graphiquement, cela se représente comme ci-dessous (Figure III.8) : le processus du score observé est représenté en rouge et les processus gaussiens simulés sont représentés en gris).
Dans le graphe (Figure III.8.a), le processus du score est inclus dans le « nuage » formé par les processus gaussiens simulés, nous pouvons voir directement que l'hypothèse nulle n'est pas rejetée. En revanche, dans le graphe (Figure III.8.b), le processus du score n'est pas toujours inclus dans le « nuage » formé par les processus gaussiens simulés, nous pouvons voir directement que l'hypothèse nulle est rejetée. Mais que dire lorsque l'on obtient le graphe (Figure III.8.c) ? Comment peut-on juger graphiquement que l'hypothèse nulle est rejetée ou pas ?
Ce type de graphe ne constitue pas un test formel ; il doit être accompagné de la p-value associée.

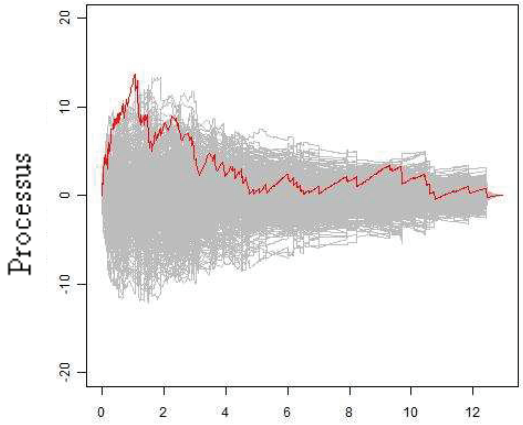
Figure III.8. Représentation graphique de l'utilisation des processus du score pour tester l'hypothèse des taux proportionnels



(a)



(b)



(c)

III.3.3.b. Résidus de martingale cumulés sur la covariable d'intérêt

Définition

Les résidus de martingales cumulées pour la covariable z_k sont définis de la manière suivante :

$$M_{cum,k}(x,\beta) = \sum_{i=1}^{n} \int_0^{\tau} I(z_{ik} \leq x) dM_i(t) \qquad \text{(III.7)}$$

Ces résidus sont un cas particulier des résidus de transformées de martingale où la fonction prévisible h_i est égale à $I(z_{ik} \leq x)$. Nous en déduisons donc que sous l'hypothèse H_0 (H_0 : « La forme fonctionnelle est correcte »), ce processus doit être d'espérance nulle avec une certaine variance.

III.3.3.b.2. Etude de la forme fonctionnelle d'une covariable

Intuition

Il a été démontré [Therneau, 1990] que de lisser les résidus de martingales en fonction de la covariable d'intérêt donnait une approximation correcte de la forme fonctionnelle de cette covariable :

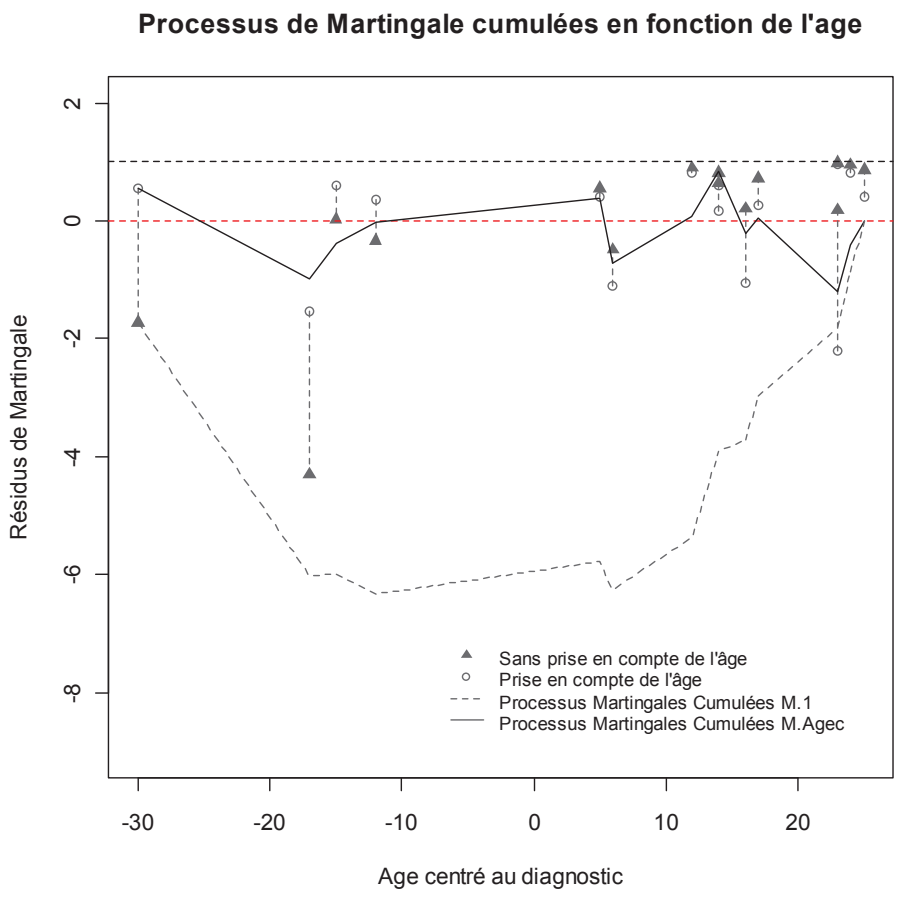
- Si la forme fonctionnelle est correcte, le lissage des résidus en fonction de la covariable d'intérêt z_k est une droite horizontale proche de zéro. Cela signifie que les résidus de martingale ont une valeur proche de zero; leur somme cumulées en fonction de la covariable d'interet doit être un processus d'espérance nulle et possédant une certaine variance.
- Si la forme fonctionnelle est incorrecte, le lissage des résidus en fonction de la covariable d'intérêt z_k n'est pas une droite horizontale proche de zéro. La forme obtenue représente une approximation de la forme fonctionnelle, cela implique qu'il existe une tendance dans la structure des résidus de martingale. Lorsque les résidus de martingale sont cumulés en fonction de la covariable d'intérêt, nous n'obtenons pas un

processus d'espérance nulle : le processus s'écarte de la valeur zéro d'une façon particulière en lien avec la tendance que représente le lissage des résidus.

En reprenant l'exemple illustré dans le paragraphe III.3.1 (figure III.1) concernant les résidus de martingale, lorsque la covariable *âge* n'est pas prise en compte dans le modèle (M.1), nous pouvons observer une tendance linéaire dans le lissage des résidus. En revanche, lorsque la covariable *âge* est incluse dans le modèle (M.Age), le lissage des résidus de martingale est proche de zéro.

Lorsque nous cumulons ces résidus de martingale en fonction de la covariable (figure III.9), nous pouvons observer que le processus obtenu après avoir ajusté le modèle M.Agec (trait plein gris foncé) est plus proche de zéro avec une variance plus faible que le processus obtenu après avoir ajusté le modèle M.1 (pointillés gris foncé) qui s'éloigne de zéro. Notons que notre échantillon étant petit, la variabilité est grande, mais ce jeu de données illustre déjà notre propos.

Figure III.9. Résidus de martingale cumulés en fonction de l'age centré



Approximation de la distribution du processus sous l'hypothèse nulle

En utilisant le développement de Taylor sur la formule (III.7), ainsi que certaines approximations présentées dans la partie III.3.3.a concernant le processus du score, il a été démontré [Lin, 1996] que la distribution de $n^{1/2}M_{cum}(x,\hat{\beta})$ était asymptotiquement équivalente à la distribution du processus W_z (Appendix 3), tel que:

$$W_z(x) = n^{1/2}M_{cum}(x,\beta_0) - n^{-1/2}U(\beta_0;\tau)\times nI(\hat{\beta};\tau)^{-1}\times J(\tau,z,\beta_0)$$

avec $J(\tau,z,\beta_0) = \sum_{i=1}^{n}\int_0^{\tau} I(z_i \le x)z_i \exp(\beta_0 z_i)d\Lambda_0(u)$, $M_{cum}(x,\beta_0)$ représentant le processus des résidus de martingale cumulées sous le vrai modèle, $U(\beta_0,\tau)$ représentant le score usuel sous le vrai modèle, $I(\beta,\tau)$ représentant la matrice information.

De la même manière que pour le processus du score, du fait que nous ne connaissons pas la distribution des temps de décès et la valeur des coefficients des covariables incluses dans le modèle sous l'hypothèse nulle, la distribution de la martingale n'est pas connue sous l'hypothèse nulle et, $U_k(\beta_0,t)$ sera approximé par $\hat{M}_{1k}(t)$:

$$\hat{M}_{1k}(t) = \sum_i \int_0^t z_{ik}dN_i(u)G_i$$

et $n^{1/2}M_{cum,k}(x,\beta_0)$ sera approximé par $\hat{P}_{1k}(x)$:

$$\hat{P}_{1k}(x) = \sum_{i=1}^{n}\int_0^{\tau} I(z_{ik} \le x)dN_i(u)G_i$$

Egalement, de la même manière que pour le processus du score, il a été montré que le processus $W_z(t)$ avait la même distribution asymptotique que le processus $\tilde{W}_z(t)$ (Appendix 2.3). La distribution des résidus de martingale cumulées peut donc être approximée par :

$$\tilde{W}_z(x) = n^{-1/2}\left(\hat{P}_1(x) - J(\tau, z, \beta_0) I\left(\hat{\beta}; \tau\right)^{-1} \hat{M}_1(\tau) \right)$$

Test statistique pour tester l'hypothèse H_0 : « La forme fonctionnelle est correcte »

Pour tester l'hypothèse H_0 : « La forme fonctionnelle est correcte » en supposant les autres hypothèses du modèle correctes, la statistique de test utilisée est la valeur maximale (en valeur absolue) du processus observé :

$$\sup_x \left| n^{-1/2} M_{cum,k}\left(x, \hat{\beta}\right) \right|$$

Le principe du test est le même que pour le processus du score : afin d'accepter ou de rejeter l'hypothèse nulle, la p-value, représentant le nombre de fois pour lesquelles la statistique de test est inférieure à la valeur maximale (en valeur absolue) des processus gaussiens simulées sera estimée.

Graphiquement, les résidus de martingale cumulés peuvent être représentés en fonction de la covariable d'intérêt z_k [Lin, 1993][Lin, 1996] :

- Pour chaque patient *i*, calculer le résidu de martingale obtenu à partir du modèle ajusté au temps T_i. Nous obtenons ainsi un vecteur de longueur égale au nombre de patients présents dans notre échantillon

$$\begin{pmatrix} \hat{M}_1(T_1,) \\ \hat{M}_2(T_2) \\ \vdots \\ \hat{M}_n(T_n) \\ \end{pmatrix}$$

- Cumuler les martingales sur la covariable d'intérêt z_k. Nous obtenons ainsi un vecteur de longueur égale à l, nombre de valeur prises par la covariable dans l'échantillon

$$\begin{pmatrix} \hat{M}_{cum,k}(z_k(1)) \\ \hat{M}_{cum,k}(z_k(2)) \\ \vdots \\ \hat{M}_{cum,k}(z_k(l)) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} \int_0^{\tau} I(z_{ik} \leq z_k(1)) d\hat{M}_i(t) \\ \sum_{i=1}^{n} \int_0^{\tau} I(z_{ik} \leq z_k(2)) d\hat{M}_i(t) \\ \vdots \\ \sum_{i=1}^{n} \int_0^{\tau} I(z_{ik} \leq z_k(l)) d\hat{M}_i(t) \end{pmatrix}$$

- Faire le graphique de la somme cumulée des résidus de martingale en fonction de la covariable d'intérêt z_k.
- Faire le graphique de plusieurs processus gaussiens correspondants en fonction de la covariable d'intérêt z_k.

Tout comme pour le processus du score, le graphe du processus des résidus de martingale cumulés en fonction de la covariable d'intérêt ne constitue pas un test formel.

III.3.3.c. Résidus de martingale cumulés sur le logarithme du taux relatif

Définition

Les résidus de martingales cumulées en fonction des valeurs du logarithme du taux relatif βz sont définis de la manière suivante :

$$M_{cum}(x, \beta) = \sum_{i=1}^{n} \int_0^{\tau} I(\beta z_i \leq x) dM_i(u) \qquad \text{(III.8)}$$

De même que précédemment, ces résidus sont un cas particulier des résidus de transformées de martingale où la fonction prévisible h_i est égale à $I(\beta z_i \leq x)$. Nous en déduisons donc que

sous l'hypothèse H_0 (H_0 : « La fonction de lien est correcte »), ce processus doit être d'espérance nulle avec une certaine variance.

Etude de la fonction de lien

Intuition

Lorsque le modèle de survie considéré n'est ajusté que sur une seule covariable, la forme fonctionnelle et la fonction de lien sont alors confondues :

$$\lambda_{obs,i}(t) = \lambda_0(t)\exp(g(z_i))$$
$$\lambda_{obs,i}(t) = \lambda_0(t)fl(z_i)$$

avec

$$fl(x) = \exp(g(x))$$

En considérant le logarithme du taux relatif βz comme étant une covariable à part entière, $\gamma = \beta z$, l'étude de la fonction de lien revient à vérifier la forme fonctionnelle utilisée pour la covariable γ. Les résidus de martingale sont donc cumulés sur la covariable γ avec la même intuition et le même raisonnement que ceux expliqués dans les parties concernant la forme fonctionnelle.

Approximation de la distribution du processus sous l'hypothèse nulle

En utilisant le développement de Taylor sur la formule (III.8), ainsi que certaines approximations présentées dans la partie III.3.3.a concernant le processus du score, il a été montré [Lin, 1996] que la distribution de $n^{1/2}M_{cum}\left(x,\hat{\beta}\right)$ était asymptotiquement équivalente à la distribution du processus W_z (Appendix 4), tel que:

$$W_z(x) = n^{1/2}M_{cum}(x,\beta_0) - n^{-1/2}U(\beta_0;\tau)\times nI\left(\hat{\beta};\tau\right)^{-1}\times J(\tau,\gamma,\beta_0)$$

avec $J(\tau,\gamma,\beta_0)=\sum_{i=1}^{n}\int_0^{\tau} I(\gamma_i \le x)\gamma_i \exp(\beta_0\gamma_i)d\Lambda_0(t)$, $M_{cum}(x,\beta_0)$ représentant le processus des résidus de martingale cumulées sous le vrai modèle, $U(\beta_0,\tau)$ représentant le score usuel sous le vrai modèle, I(β,τ) représentant la matrice d'information.

De la même manière que pour les tests précédents, du fait que nous ne connaissons pas la distribution des temps de décès et la valeur des coefficients des covariables incluses dans le modèle sous l'hypothèse nulle, la distribution de la martingale n'est pas connue sous l'hypothèse nulle et, $U_k(\beta_0,t)$ sera approximé par $\hat{M}_{1k}(t)$:

$$\hat{M}_{1k}(t)\approx\sum_i\int_0^t z_{ik}\,dN_i(u)G_i$$

et $n^{1/2}M_{cum}(x,\beta_0)$ sera approximée par $\hat{P}_1(x)$:

$$\hat{P}_1(x)=\sum_{i=1}^{n}\int_0^{\tau} I(\hat{\gamma}_i \le x)dN_i(t)G_i$$

De la même manière que précédemment (Appendix 2.3), il a été montré que les processus $\tilde{W}_z(t)$ et $W_z(t)$ avaient la même distribution asymptotiquement. La distribution des résidus de martingale cumulé peut donc être approximée par :

$$\tilde{W}_z(x)\approx n^{-1/2}\left(\hat{P}_1(x)-J(\tau,\gamma,\beta_0)I\left(\hat{\beta};\tau\right)^{-1}\hat{M}_1(\tau)\right)$$

Test statistique pour tester l'hypothèse H_0 : « La fonction de lien est correcte »

Pour tester l'hypothèse H_0 : « La fonction de lien est correcte » en supposant les autres hypothèses du modèle correctes, un test formel peut être effectué avec visualisation graphique de la même manière qu'est présenté le test pour la forme fonctionnelle pour une covariable.

III.4 Développement d'une « boîte à outils » permettant de tester les hypothèses d'un modèle paramétrique du taux en excès

Les modèles de taux en excès considérés dans cette partie reposent sur la décomposition suivante :

$$\lambda_o(t,z) = \lambda_c(t,z) + \lambda_a(a+t,x)$$

λ_o représentant le taux de mortalité observé au temps t, λ_c représentant le taux de mortalité en excès au temps t, λ_a représentant le taux de mortalité attendu au temps t, x représentant le vecteur des covariables démographiques et z représentant le vecteur des variables pronostiques incluses dans le modèle.

Les résidus développés pour des modèles de taux en excès reposent également sur la théorie des martingales comme exposé dans le cadre de la survie globale.
Les processus de martingale dans la cadre de la survie nette sont semblables aux processus de martingale dans le cadre de la survie globale. Ils s'expriment de la manière suivante :

$$\begin{aligned}
M_i(t) &= N_i(t) - \int_0^t Y_i(u)\lambda_o(u,z_i)du \\
M_i(t) &= N_i(t) - \int_0^t Y_i(u)[\lambda_0(u)\exp(\beta z_i) + \lambda_a(a_i+u,x_i)]du \\
M_i(t) &= N_i(t) - \int_0^t Y_i(u)\lambda_0(u)\exp(\beta z_i)du - \int_0^t Y_i(u)\lambda_a(a_i+u,x_i)du \\
M_i(t) &= \widetilde{N}_i(t) - \int_0^t Y_i(u)\lambda_0(u)\exp(\beta z_i)du
\end{aligned}$$

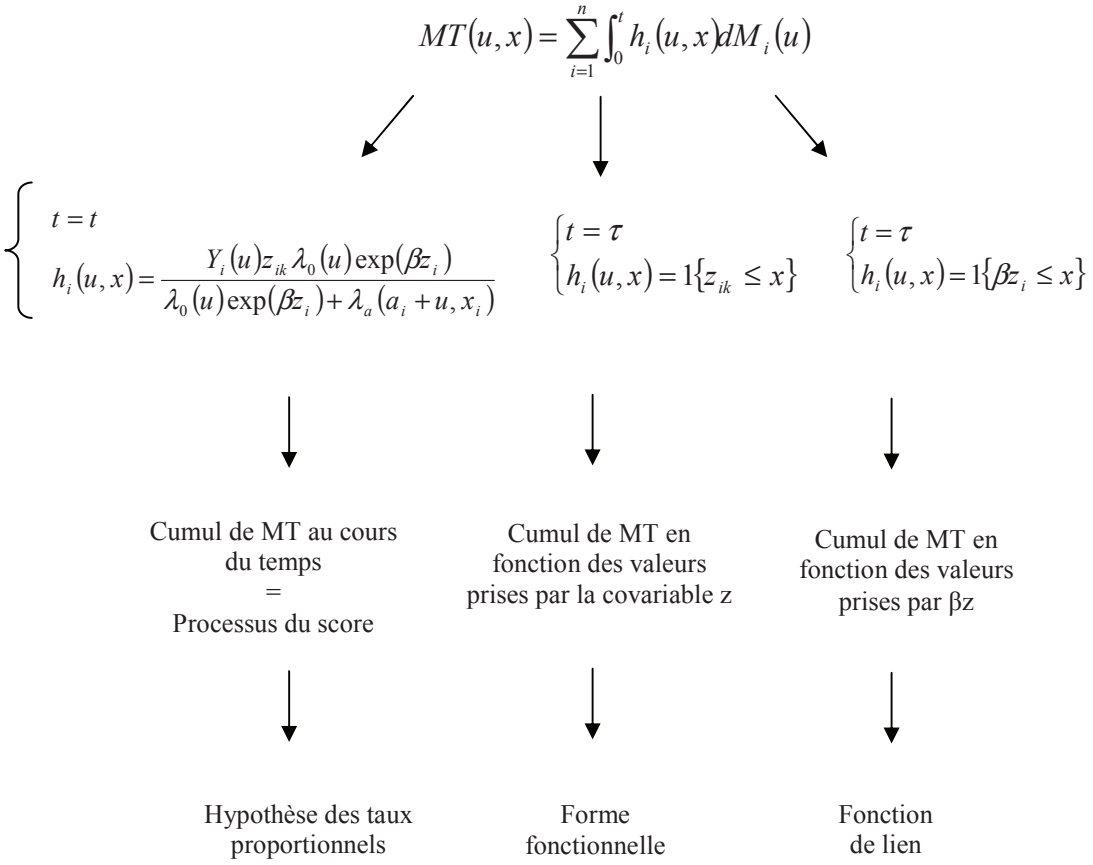
$M_i(t)$ représentant la martingale pour le patient i au temps t, le premier terme de droite $\widetilde{N}_i(t)$ représentant le processus de comptage compensé pour le patient i au temps t et le second terme de droite représentant le compensateur du processus de comptage compensé au temps t.

Contrairement au cadre de la survie globale (notamment pour le modèle de Cox), très peu d'outils diagnostiques de l'adéquation d'un modèle aux données ont été développés dans

le cadre des modèles paramétriques du taux en excès. Une seule méthode [Stare, 2005] disponible et utilisable sous le logiciel R, a été développée de manière analogue aux résidus de Schoenfeld [Schoenfeld, 1982] pour les modèles paramétriques du taux en excès. Cette méthode permet de tester l'hypothèse des taux proportionnels à l'aide d'un test formel. Un autre test, reposant sur le processus du score pour les modèles du taux en excès semi-paramétrique [Cortese, 2008] a été développé de manière analogue à celui de Lin pour le modèle de Cox. Cette méthode ne sera pas abordée dans la suite du fait que nous ne nous intéressons qu'aux modèles paramétriques du taux en excès (il faut noter de plus que, d'un point de vue pratique, le package R qui implémente cette méthode, présente de nombreux problèmes de fonctionnement ; nous n'avons ainsi pas réussi à obtenir des résultats cohérents avec ce package).

L'objectif de cette partie de thèse est de développer une boîte à outils composée de différents tests formels, issus du même cadre théorique, celui des transformées de martingale. Ils devront permettre, dans le cadre du taux de mortalité en excès, de tester l'hypothèse de proportionnalité des taux, la forme fonctionnelle associée à une covariable ainsi que la fonction de lien. Pour cela, il s'est agit d'adapter les méthodes décrites dans le cadre de la survie globale (partie III.3), développées par Lin [Lin, 1996], au contexte de la survie nette. La démarche intellectuelle a été identique à celle de Lin, ce qui a mené pour certains cas, a considérer une autre fonction prévisible h_i (Figure III.10). Comme il le sera indiqué, l'adaptation en survie nette repose sur des développements décrits dans le cadre de la survie globale (partie III.3) qui ont pu être transposés au cadre de la survie nette. Ces développements ne seront pas toujours re-détaillés.

Figure III.10. Trois choix de fonction h_i pour définir trois transformées de martingale différentes pour tester les trois principales hypohèses dans le cadre de la survie nette



III.4.1 Résidus de Stare

Définition

Les résidus introduits par Stare et al [Stare, 2005] sont des résidus semblables aux résidus de Schoenfeld dans le cadre des modèles de taux en excès, construits pour tester l'hypothèse de proportionnalité. Ils s'expriment de la manière suivante (pour la covariable *k*) :

$$U_{ik}(\beta) = z_{ik} - \frac{\sum_{j \in R_i} z_{jk} \left\{ \lambda_a (a_j + t_i, x_j) + \lambda_0 (t_i) \exp(\beta z_j) \right\}}{\sum_{j \in R_i} \left\{ \lambda_a (a_j + t_i, x_j) + \lambda_0 (t_i) \exp(\beta z_j) \right\}}$$

avec z_{ik} représentant la valeur de la k^{ème} covariable pronostique du patient décédant au i^{ème} temps de décès, z_j représentant le vecteur de covariables du patient j encore à risque au temps t_i, λ_a étant le taux de mortalité attendus, λ_0 le taux de mortalité de base en excès, x_j étant le vecteur de covariables démographiques du patient *j*, et β étant le vecteur des coefficients des covariables incluses dans le modèle.

Tout comme pour les résidus de Schoenfeld construits à partir du modèle de Cox (partie III.3.2), on a :

$$U_{ik}(\beta) = z_{ik} - \bar{z}(\beta, t_i)$$

Ces résidus représentent la différence entre la valeur observée de la $k^{ième}$ covariable de l'individu décédant au $i^{ième}$ temps et la valeur attendue de cette covariable pour le patient décédé sous le modèle proportionnel ajusté. Cette valeur attendue est une moyenne pondérée par le risque de décès des patients encore à risque ; il s'agit du profil moyen des individus encore à risque au temps t_i.

Test de l'hypothèse des taux proportionnels

Intuition

D'un point de vue graphique, ces résidus étant analogues aux résidus de Schoenfeld dans le cadre de la survie nette, le lissage des résidus standardisé en fonction du temps fournit une approximation de l'effet de la covariable d'intérêt en fonction du temps [Stare, 2005] :

- Si l'effet est proportionnel, le lissage des résidus standardisés en fonction du temps est une droite horizontale proche de zéro.
- Si l'effet est non-proportionnel, le lissage des résidus standardisés en fonction du temps représente une approximation de l'effet de la covariable d'intérêt ; cela implique qu'il existe une tendance dans la structure des résidus de Stare.

Si l'effet est proportionnel, le cumul de ces résidus standardisé oscille autour de zéro tout au long du suivi avec une certaine variance. En revanche, si l'effet est non-proportionnel, le processus s'éloigne de zéro.

Approximation de la distribution du processus sous l'hypothèse nulle

Stare introduit tout d'abord les résidus standardisés R_i tels que :

$$R_i(\beta_0) = \frac{U_i(\beta_0)}{\sqrt{V_i(\beta_0)}} \quad \text{avec} \quad V_i \text{ représentant la variance du résidu } U_i$$

En les cumulant et en changeant l'échelle des temps, nous obtenons le processus suivant :

$$\begin{cases} B_n\left(\beta_0, \frac{k}{n}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{k} R_i(\beta_0) \\ B(\beta_0, 0) = 0 \end{cases}$$

avec $k = 1,\ldots,n$ représentant le nombre de décès

$i = 1,\ldots,k$ représentant l'indice des résidus participant à la somme cumulée

k/n représentant le temps du $k^{\text{ième}}$ décès dans une nouvelle échelle de temps permettant d'avoir des temps de décès équidistant dans l'intervalle [0 ;1]

Cette quantité ayant les propriétés d'un mouvement brownien [Stare, 2005], elle amène le processus

$$BP_n\left(\beta_0, \frac{k}{n}\right) = \frac{1}{\sqrt{n}}\left\{\sum_{i=1}^{k} R_i(\beta_0) - \frac{k}{n}\sum_{i=1}^{n} R_i(\beta_0)\right\} \qquad \text{pour} \quad k = 1,\ldots,n$$

à converger vers un pont brownien.

En pratique, β_0 n'est pas connu, sa valeur est donc remplacée par la valeur estimée $\hat{\beta}$. En utilisant cette approximation les résidus ne sont pas iid, les propriétés du processus B_n changent; il n'a alors plus les caractéristiques d'un mouvement brownien. Il a cependant été démontré que la distribution du processus $BP_n\left(\hat{\beta}, \frac{k}{n}\right)$ pouvait être approximée asymptotiquement par un pont brownien [Pohar-Perme, 2007].

Test statistique pour tester l'hypothèse H_0 : « L'hypothèse des taux proportionnels est correcte »

La statistique de test utilisée pour tester l'hypothèse H_0 est la valeur maximale (en valeur absolue) du processus observé :

$$\max\left|BP_n\left(\hat{\beta}, \frac{k}{n}\right)\right|$$

D'après le théorème de Smirnov et Kolmogorov, la distribution du maximum de la valeur absolue d'un pont brownien BB est la suivante :

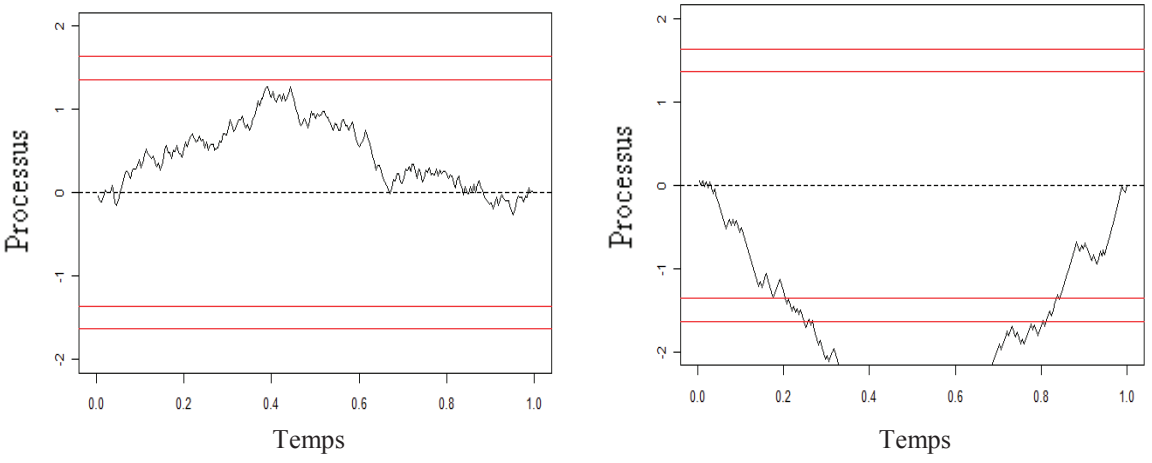
$$P\left(\max_{u\in[0;1]}\left(\left|BB(u)\right|\right) \le x\right) = 1 + 2\sum_{k=1}^{\infty}(-1)^{k} e^{-2k^2x^2} \quad \text{avec} \quad x > 0$$

Si le maximum de la valeur absolue du pont brownien est inférieur à la valeur critique correspondant à un α à 5%, soit une valeur critique de 1.361, le modèle est considéré comme correct. Autrement, le modèle est considéré comme incorrect.
Contrairement aux tests de Lin [Lin, 1996] présentés dans le cadre de la survie globale, ce test ne nécessite pas de simulation de processus suivant la distribution du processus observé sous l'hypothèse H_0.

Graphiquement, le pont brownien obtenu à partir des résidus doit être représenté en fonction du temps: si le pont brownien ne dépasse pas la valeur critique des 95%, l'hypothèse nulle est acceptée, autrement, elle est rejetée.

Voici un exemple de représentation graphique de pont brownien pour une covariable à effet proportionnel (Figure III.11.a) et à effet non-proportionnel (Figure III.11.b).
Le pont brownien est représenté en noir et les valeurs critiques à 95% et 99% sont représentées en rouge.

Figure III.11. Représentation des résidus de Stare en fonction du temps



(a) (b)

Contrairement aux différents tests de Lin présentés précédemment, le graphe de Stare nous permet de voir directement le résultat du test, c'est-à-dire, nous pouvons directement voir si l'hypothèse nulle est rejetée ou pas.

III.4.2 Proposition d'un test reposant sur le processus du score pour tester l'hypothèse de proportionnalité

La démarche intellectuelle de cette sous-partie a été d'écrire le score, défini à partir de la vraisemblance, comme une transformée de martingale en définissant son propre processus prévisible. Cette sous-partie nécessite de se rapporter aux développements effectués pour l'étude du processus du score dans le cadre de la survie globale ainsi qu'aux Appendix 2 et 5. Pour alléger l'écriture, nous supposerons qu'il n'y a qu'une seule covariable incluse dans le modèle.

Définition

Par définition, la log-vraisemblance d'un modèle paramétrique de taux en excès est la suivante :

$$L_i(\beta) = -\int_0^{t_i} \lambda_c(u, z_i)du - \int_0^{t_i} \lambda_a(a_i + u, x_i)du + \delta_i \times \log[\lambda_c(t_i, z_i) + \lambda_a(a_i + t_i, x_i)]$$

Le score s'écrit alors de la manière suivante (Appendix 5) :

$$U(\beta) = \frac{\partial}{\partial \beta} L(\beta)$$

$$U(\beta) = \sum_i \left[-\int_0^{t_i} z_i \lambda_0(u) \exp(\beta z_i)du + \delta_i \times \frac{z_i \lambda_0(t_i) \exp(\beta z_i)}{\lambda_0(t_i) \exp(\beta z_i) + \lambda_a(a_i + t_i, x_i)} \right]$$

En utilisant les processus de comptage et en posant τ la fin du suivi, le score devient (Appendix 5) :

$$U(\beta) = \sum_i \left[-\int_0^{\tau} Y_i(u) z_i \lambda_0(u) \exp(\beta z_i)du + \int_0^{\tau} \frac{Y_i(u) z_i \lambda_0(u) \exp(\beta z_i)}{\lambda_0(t_i) \exp(\beta z_i) + \lambda_a(a_i + u, x_i)} dN_i(u) \right]$$

Le processus du score au temps *t* peut alors s'écrire de la manière suivante (Appendix 5) :

$$U(\beta,t)=\sum_i\left[-\int_0^t Y_i(u)z_i\lambda_0(u)\exp(\beta z_i)du+\int_0^t\frac{Y_i(u)z_i\lambda_0(u)\exp(\beta z_i)}{\lambda_0(t_i)\exp(\beta z_i)+\lambda_a(a_i+u,x_i)}dN_i(u)\right]$$

Ce qui équivaut à l'écrire comme une transformée de martingale (Appendix 5) :

$$U(\beta,t)=\sum_i\left[\int_0^t\frac{z_i\lambda_0(u)\exp(\beta z_i)}{\lambda_0(t_i)\exp(\beta z_i)+\lambda_a(a_i+u,x_i)}dM_i(u)\right]$$

De même que dans le cadre de la survie globale, écrire le processus du score comme une transformée de martingale permet de déduire sa distribution sous l'hypothèse nulle : le processus doit fluctuer autour de zéro avec une variance égale à $E(< U >_t)$.

Test de l'hypothèse des taux proportionnels

Intuition

L'idée intuitive est similaire à celle exposée dans le cadre de la survie globale.

Approximation de la distribution du processus du score sous l'hypothèse nulle

Les approximations utilisées dans le cadre de la survie globale peuvent être utilisée dans le cadre de la survie nette (Appendix 2.2/Appendix 2.3). En effet, en passant au cadre de la survie nette, seule le processus prévisible h_i change :

$$h_i(t)=\frac{z_i\lambda_0(t)\exp(\beta z_i)}{\lambda_0(t)\exp(\beta z_i)+\lambda_a(a_i+t,x_i)}$$

En utilisant le développement de Taylor pour la variable β (Appendix 2.2),

$$W_z(t)=n^{-1/2}\left(U(\beta_0;t)-I(\hat{\beta};t)I(\hat{\beta};\tau)^{-1}U(\beta_0;\tau)\right)$$

$U(\beta_0,t)$ représentant le processus du score sous l'hypothèse nulle au temps t, $U(\beta_0,\tau)$ représentant le processus du score sous l'hypothèse nulle au temps τ (score usuel), $I(\beta,t)$ représentant l'opposé de la dérivée seconde du logarithme de la vraisemblance au temps t et $I(\beta,\tau)$ représentant l'opposé de la dérivée seconde du logarithme de la vraisemblance au temps τ (matrice information).

De la même manière que pour la survie globale, du fait que nous ne connaissons pas la distribution des temps de décès et la valeur des coefficients des covariables incluses dans le modèle sous l'hypothèse nulle, le vecteur β_0 sera remplacé par son estimateur $\hat{\beta}$ et la martingale M_i sera approximée par $N_i(u).G_i$ (G_i suivant la distribution d'une loi normale standard). La distribution de $n^{-1/2}U\left(\hat{\beta},t\right)$ est asymptotiquement équivalente à la distribution du processus $\tilde{W}_z(t,z)$ (Appendix 2.3), tel que :

$$\tilde{W}_z(t,z) = n^{-1/2}\left(\hat{M}_1(t) - I\left(\hat{\beta};t\right)I\left(\hat{\beta};\tau\right)^{-1}\hat{M}_1(\tau)\right)$$

avec

$$\hat{M}_1(t) = \sum_i \int_0^t \frac{z_i Y_i(u)\lambda_0(u)\exp\left(\hat{\beta}z_i\right)}{\lambda_0(u)\exp\left(\hat{\beta}z_i\right) + \lambda_a(a_i+u,x_i)} dN_i(u)G_i$$

Test statistique pour tester l'hypothèse H_0 : « L'hypothèse des taux proportionnels est correcte »

Pour tester l'hypothèse H_0 : « L'hypothèse des taux proportionnels est correcte » en supposant les autres hypothèses du modèle correctes, la statistique de test utilisée est la valeur maximale (en valeur absolue) du processus du score observé :

$$\sup_t \left| n^{1/2}U\left(\hat{\beta},t\right) \right|$$

Sous l'hypothèse nulle, la distribution du processus $n^{1/2}U(\hat{\beta},t)$ est asymptotiquement équivalente à la distribution d'un processus gaussien particulier comme vu précédemment.
Afin d'accepter ou de rejeter l'hypothèse nulle, la p-value, représentant le nombre de fois pour lesquelles la statistique de test est inférieure à la valeur maximale (en valeur absolue) des processus gaussiens simulées sera estimée :

$$p.val = \frac{\sum_i \left(\sup_x \left| n^{1/2} U_k(\hat{\beta},t) \right| < \sup |sim_i| \right)}{n.sim}$$

Avec *n.sim* représentant le nombre de processus gaussiens simulés et $|sim_i|$ représentant le processus gaussien *i* en valeur absolue.

Autrement dit, en fixant α à 5%, l'hypothèse de proportionnalité sera rejetée si la statistique de test est supérieure au maximum des processus simulés dans plus de 95% des cas.

III.4.3 Proposition d'un test basé sur les résidus de martingale cumulés sur la covariable d'intérêt pour tester la forme fonctionnelle d'une covariable

Therneau [Therneau, 1990] a montré que dans le cadre de la survie globale, l'espérance des résidus de martingale était proportionnelle à la forme fonctionnelle des covariables. En supposant que ceci peut être appliqué au cadre de la survie nette, la même fonction prévisible h_i utilisée pour tester la forme fonctionnelle en survie globale peut être utilisée.
Cette sous-partie nécessite de se rapporter aux développements effectués pour l'étude des résidus de martingale cumulés pour tester la forme fonctionnelle dans le cadre de la survie globale ainsi qu'aux Appendix 2, 3 et 6.
De même que précédemment, pour alléger l'écriture, nous supposerons qu'il n'y a qu'une seule covariable incluse dans le modèle.

Définition

Les résidus de martingales cumulées pour la covariable z s'expriment de la manière suivante :

$$M_{cum}(x,\beta) = \sum_{i=1}^{n} \int_0^{\tau} I(z_i \le x) dM_i(u)$$

avec $M_i(t) = N_i(t) - \int_0^t Y_i(u)\lambda_0(u)\exp(\beta z_i)du - \int_0^t Y_i(u)\lambda_a(a_i + u, x_i)du$

Etude de la forme fonctionnelle d'une covariable

L'idée intuitive est similaire à celle exposée dans le cadre de la survie globale.

De même que pour les processus du score, les approximations utilisées dans le cadre de la survie globale peuvent être utilisée dans le cadre de la survie nette (Appendix 6). Le processus est similaire que l'on se place dans le cadre de la survie globale ou dans le cadre de la survie nette. Le changement réside dans l'expression des processus gaussiens simulés approximant la distribution du processus sous l'hypothèse nulle.

En utilisant donc les mêmes approximations que dans le cadre de la survie globale, nous pouvons affirmer que la distribution de $n^{1/2} M_{cum}\left(x, \hat{\beta}\right)$ est asymptotiquement équivalente à la distribution du processus W_z, tel que:

$$W_z(x) = n^{1/2} M_{cum}(x, \beta_0) - n^{-1/2} U(\beta_0; \tau) \times nI\left(\hat{\beta}; \tau\right)^{-1} \times J(\tau, z, \beta_0)$$

avec $J(\tau, z, \beta_0) = \sum_{i=1}^{n} \int_0^{\tau} I(z_i \leq x) z_i \exp(\beta_0 z_i) d\Lambda_0(t)$, $M_{cum}(x, \beta_0)$ représentant le processus des résidus de martingale cumulées sous le vrai modèle, $U(\beta_0,\tau)$ représentant le score usuel sous le vrai modèle, $I(\beta,\tau)$ représentant la matrice information.

Du fait que la distribution de la martingale ne soit pas connue sous l'hypothèse nulle, comme nous l'avons vu précédemment, $U(\beta_0;t)$ sera approximé par $\hat{M}_1(t)$:

$$\hat{M}_1(t) = \sum_i \int_0^t \frac{z_i \lambda_0(u) \exp\left(\hat{\beta} z_i\right)}{\lambda_0(u) \exp\left(\hat{\beta} z_i\right) + \lambda_e(a_i + u, x_i)} dN_i(u) G_i$$

et $n^{1/2} M_{cum}(x, \beta_0)$ sera approximée par $\hat{P}_1(x)$:

$$\hat{P}_1(x) = \sum_{i=1}^{n} \int_0^{\tau} I(z_i \leq x) dN_i(t) G_i$$

La distribution des résidus de martingale cumulés peut donc être approximée par le processus $\tilde{W}_z(t)$ qui a la même distribution asymptotique que le processus $W_z(t)$ (Appendix 2.3) :

$$\tilde{W}_z(x) = n^{-1/2} \left(\hat{P}_1(x) - J(\tau, z, \beta_0) I\left(\hat{\beta}; \tau\right)^{-1} \hat{M}_1(\tau) \right)$$

Afin d'étudier la forme fonctionnelle de la covariable d'intérêt, un test formel peut être effectué avec visualisation graphique de la même manière que décrite dans le cadre de la survie globale.

III.4.4 Proposition d'un test basé sur les résidus de martingale cumulés sur le logarithme du taux relatif pour tester la fonction de lien

Pour les mêmes raisons qu'exposées pour le test concernant la forme fonctionnelle, la même fonction prévisible h_i utilisée pour tester la fonction de lien en survie globale peut être utilisée. Cette sous-partie nécessite de se rapporter aux développements effectués pour l'étude des résidus de martingale cumulés pour tester la forme fonctionnelle dans le cadre de la survie globale ainsi qu'aux Appendix 2, 4 et 6.

Définition

Les résidus de martingales cumulées en fonction des valeurs du log du taux relatif βz s'expriment de la manière suivante :

$$M_{cum,k}(x,\beta) = \sum_{i=1}^{n} \int_0^{\tau} I(\beta z_i \le x) dM_i(u)$$

avec $M_i(t) = N_i(t) - \int_0^t Y_i(u)\lambda_0(u)\exp(\beta z_i)du - \int_0^t Y_i(u)\lambda_a(a_i+u, x_i)du$

Etude de la fonction de lien

L'idée intuitive est similaire à celle exposée dans le cadre de la survie globale.

En utilisant les mêmes approximations que dans le cadre de la survie globale, nous pouvons affirmer que la distribution de $n^{1/2} M_{cum}(x,\hat{\beta})$ est asymptotiquement équivalente à la distribution du processus W_z (Appendix 4/Appendix 6), tel que:

$$W_z(x) = n^{1/2} M_{cum}(x,\beta_0) - n^{-1/2} U(\beta_0;\tau) \times nI(\hat{\beta};\tau)^{-1} \times J(\tau,\gamma,\beta_0)$$

Avec $J(\tau,\gamma,\beta_0)=\sum_{i=1}^{n}\int_0^{\tau} I(\gamma_i \le x)\gamma_i \exp(\beta_0\gamma_i)d\Lambda_0(t)$, $M_{cum}(x,\beta_0)$ représentant le processus des résidus de martingale cumulées sous le vrai modèle, $U(\beta_0,\tau)$ représentant le score usuel sous le vrai modèle, $I(\beta,\tau)$ représentant la matrice information.

Du fait que la distribution de la martingale ne soit pas connue sous l'hypothèse nulle, comme précédemment, $U(\beta_0; t)$ sera approximé par $\hat{M}_1(t)$:

$$\hat{M}_1(t)=\sum_i\int_0^t \frac{z_i\lambda_0(u)\exp(\hat{\beta}z_i)}{\lambda_0(u)\exp(\hat{\beta}z_i)+\lambda_a(a_i+u,x_i)}dN_i(u)G_i$$

et $n^{1/2}M_{cum}(x,\beta_0)$ sera approximée par $\hat{P}_1(x)$:

$$\hat{P}_1(x)=\sum_{i=1}^{n}\int_0^{\tau} I(\hat{\gamma}_i \le x)dN_i(t)G_i$$

Les processus $\tilde{W}_z(t)$ et $W_z(t)$ ayant la même distribution asymptotique (Appendix 2.3), la distribution des résidus de martingale cumulées peut donc être approximée par :

$$\tilde{W}_z(x)=n^{-1/2}\left(\hat{P}_1(x)-J(\tau,z,\beta_0)I(\hat{\beta};\tau)^{-1}\hat{M}_1(\tau)\right)$$

Afin d'étudier la fonction de lien du modèle, un test formel peut être effectué avec visualisation graphique de la même manière qu'est fait le test pour la forme fonctionnelle pour une covariable décrite dans le cadre de la survie globale et dans le cadre de la survie nette.

III.4.5. Récapitulatif des différents résidus et leur relation

Résidus reposant sur les processus du score

Le tableau III.2 va nous permettre de voir les relations qui existent entre les différentes quantités reposant directement sur le processus du score.

Notons que nous utilisons ici le terme « processus » de façon un peu abusive puisque rigoureusement un processus est une famille de variable aléatoire indexée par t, alors que nous l'utilisons aussi pour désigner ces variables aléatoires.

Les lignes représentent chaque patient et les colonnes représentent les différents temps de suivi (t_1, …, t_m), avec $t_m = \tau$, temps maximal de suivi.
Chaque case, $U_i(\beta,t_j)$, représente alors le processus du score individuel du patient i au temps t_j, j = {1,…,m}. Nous supposerons que les patients sont ordonnés par ordre croissant de temps de suivi et que les patients 2 et 4 sont censurés (note : si $t > t_i$, $U_i(\beta,t) = U_i(\beta,t_i)$).

Le processus du score au temps t, $U(\beta,t)$, (gris intermédiaire entouré de bleu), représente la somme des processus du score individuels $U_i(\beta,t)$ au temps t, c'est-à-dire, la somme de la dérivée de la log-vraisemblance individuelle par rapport au coefficient de la covariable d'intérêt.
Le score, comme on l'entend de manière usuelle, représente la valeur du processus du score en $\tau = t_m$, $U(\beta,\tau) = U(\beta,t_m)$, temps maximal de suivi.
Le résidu de Schoenfeld (gris foncé entouré vert) au temps t_i représente le score partiel individuel du patient décédé en t_i. (ils ne sont considérés que pour des modèles semi-paramétriques).
Note : Le processus du score n'est pas à confondre avec les résidus du score. Le résidu du score d'un individu i (gris clair entouré de rouge) représente le score individuel du patient i (c'est-à-dire, le processus du score individuel du patient i au temps T_i).

Table III.2. Quantités reposant sur les processus du score

			Processus du score individuel à chaque temps						
Patient i	$L_i(\beta)$	$dL_i(\beta)/d\beta$	$U_i(\beta,t_1)$	$U_i(\beta,t_2)$	$U_i(\beta,t_3)$	$U_i(\beta,t_4)$	$U_i(\beta,t_5)$	.	$U_i(\beta,t_m)$
1	$L_1(\beta)$	$dL_1(\beta)/d\beta$	$U_1(\beta,t_1)$	$U_1(\beta,t_2)$	$U_1(\beta,t_3)$	$U_1(\beta,t_4)$	$U_1(\beta,t_5)$	.	$U_1(\beta,t_m)$
2	$L_2(\beta)$	$dL_2(\beta)/d\beta$	$U_2(\beta,t_1)$	$U_2(\beta,t_2)$	$U_2(\beta,t_3)$	$U_2(\beta,t_4)$	$U_2(\beta,t_5)$	.	$U_2(\beta,t_m)$
3	$L_3(\beta)$	$dL_3(\beta)/d\beta$	$U_3(\beta,t_1)$	$U_3(\beta,t_2)$	$U_3(\beta,t_3)$	$U_3(\beta,t_4)$	$U_3(\beta,t_5)$	.	$U_3(\beta,t_m)$
4	$L_4(\beta)$	$dL_4(\beta)/d\beta$	$U_4(\beta,t_1)$	$U_4(\beta,t_2)$	$U_4(\beta,t_3)$	$U_4(\beta,t_4)$	$U_4(\beta,t_5)$	.	$U_4(\beta,t_m)$
5	$L_5(\beta)$	$dL_5(\beta)/d\beta$	$U_5(\beta,t_1)$	$U_5(\beta,t_2)$	$U_5(\beta,t_3)$	$U_5(\beta,t_4)$	$U_5(\beta,t_5)$	.	$U_5(\beta,t_m)$
	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
n	$L_n(\beta)$	$dL_n(\beta)/d\beta$	$U_n(\beta,t_1)$	$U_n(\beta,t_2)$	$U_n(\beta,t_3)$	$U_n(\beta,t_4)$	$U_n(\beta,t_5)$	.	$U_n(\beta,t_m)$
Somme	$L(\beta)$	$dL(\beta)/d\beta$	$U(\beta,t_1)$	$U(\beta,t_2)$	$U(\beta,t_3)$	$U(\beta,t_4)$	$U(\beta,t_5)$	.	$U(\beta,t_m)$

Légende :

Résidus du score

Processus du score

Résidus de Schoenfeld

Notation :

$L_i(\beta)$: Contribution à la log-vraisemblance du patient i au temps τ, temps de suivi maximum.

$L(\beta)$: Log-vraisemblance au temps τ

$U_i(\beta,t_j)$: Processus du score du patient *i* au temps t_j

$U(\beta,t_j)$: Processus du score aux temps t_j

$dL_i(\beta)/d\beta = U_i(\beta,\tau)$: Contribution au score du patient i au temps τ

$dL(\beta)/d\beta = U(\beta,\tau)$: Score sur $[0 ; \tau]$ = nul théoriquement

Résidus reposant sur les martingales

Le tableau III.3 va nous permettre de voir les relations qui existent entre les différentes quantités reposant directement sur les martingales (sans transformation).

Les lignes représentent chaque patient et les colonnes représentent les différents temps de suivi (t_1, …, t_m), avec $t_m = \tau$, temps maximal de suivi.
Chaque case représente alors le processus de martingale individuel du patient i au temps t_j, j = {1,…,m}. Nous supposerons que les patients sont ordonnés par ordre croissant de la covariable z, covariable d'intérêt, t_i étant le temps de décès du patient i.
Les différentes cellules associées au patient i ayant comme temps de suivi t_i et comme valeur de covariable $z_i = z_1$ ou z_2…, représentent les martingales individuelles du patient i à chaque temps d'intérêt (note : si $t > t_i$, $M_i(t,z_i) = M_i(t_i,z_i)$).

Table III.3. Quantités reposant sur les martingales

	Processus de martingale individuel pour chaque temps de suivi				Résidu de martingale cumulés
Patient i	$M_i(t_1,z_i)$	$M_i(t_2,z_i)$	$M_i(t_3,z_i)$	$M_i(t_m,z_i)$	
1	$M_1(t_1,z_1)$	$M_1(t_2,z_1)$	$M_1(t_3,z_1)$	$M_1(t_m,z_1)$	$M_1(t_m,z_1)$
2	$M_2(t_1,z_2)$	$M_2(t_2,z_2)$	$M_2(t_3,z_2)$	$M_2(t_m,z_2)$	$M_1(t_m,z_1) + M_2(t_m,z_2)$
3	$M_3(t_1,z_3)$	$M_3(t_2,z_3)$	$M_3(t_3,z_3)$	$M_3(t_m,z_3)$	$M_1(t_m,z_1) + M_2(t_m,z_2)+ M_3(t_m,z_3)$
4	$M_4(t_1,z_4)$	$M_4(t_2,z_4)$	$M_4(t_3,z_4)$	$M_4(t_m,z_4)$	.
5	$M_5(t_1,z_5)$	$M_5(t_2,z_5)$	$M_5(t_3,z_5)$	$M_5(t_m,z_5)$	.
.	.	.	.	.	.
.	.	.	.	.	.
n	$M_n(t_1,z_n)$	$M_n(t_2,z_n)$	$M_n(t_3,z_n)$	$M_n(t_m,z_n)$	$M_1(t_m,z_1) +…+ M_n(t_m,z_n)$
Somme	$M(t_1)$	$M(t_2)$	$M(t_3)$	$M(t_m)$	

Légende :

Processus de martingale

Résidus de Martingale

Résidus de martingales cumulés

Les résidus de martingale (gris clair entourés de vert) représentent pour l'individu *i*, la différence entre l'évènement observé et le taux de mortalité prédit par le modèle au temps t_i, temps de suivi du patient *i*.

Les résidus de martingale cumulés en fonction de la variable d'intérêt (gris foncé entourés rouge) correspondent au cumul des résidus de martingale ordonnés par valeurs de la variable d'intérêt (covariable d'intérêt pour le cas de l'étude de la forme fonctionnelle ou logarithme du taux relatif pour le cas de l'étude de la fonction de lien).

Il faut noter que le processus de martingale au temps *t* (gris intermédiaire entourés de bleu), qui n'est pas utilisé, correspond à la somme des martingales de chaque individu au temps *t*, c'est-à-dire, à la somme de la différence entre l'évènement observé au temps *t* pour le patient *i* et le taux de mortalité prédit par le modèle au temps *t* pour le patient *i*.

Schéma récapitulatif des résidus en survie globale et en survie nette

Les figures III.12 à III.15 nous montrent de quelle manière chacun des résidus étudié dans ce chapitre peut s'écrire à partir de la formule générale de la transformée de martingale, $\int hdM$.

En ce qui concerne la survie globale, il s'agira des résidus de martingale, des résidus de Schoenfeld, du processus du score, des résidus de martingale cumulés en fonction de la covariable d'intérêt et des résidus de martingale cumulés en fonction du logarithme du taux relatif.

En ce qui concerne la survie nette, il s'agira des résidus de Stare, du processus du score, des résidus de martingale cumulés en fonction de la covariable d'intérêt et des résidus de martingale cumulés en fonction du logarithme du taux relatif.

Figure III.12. Résidus calculés individuellement dans le cadre de la survie globale

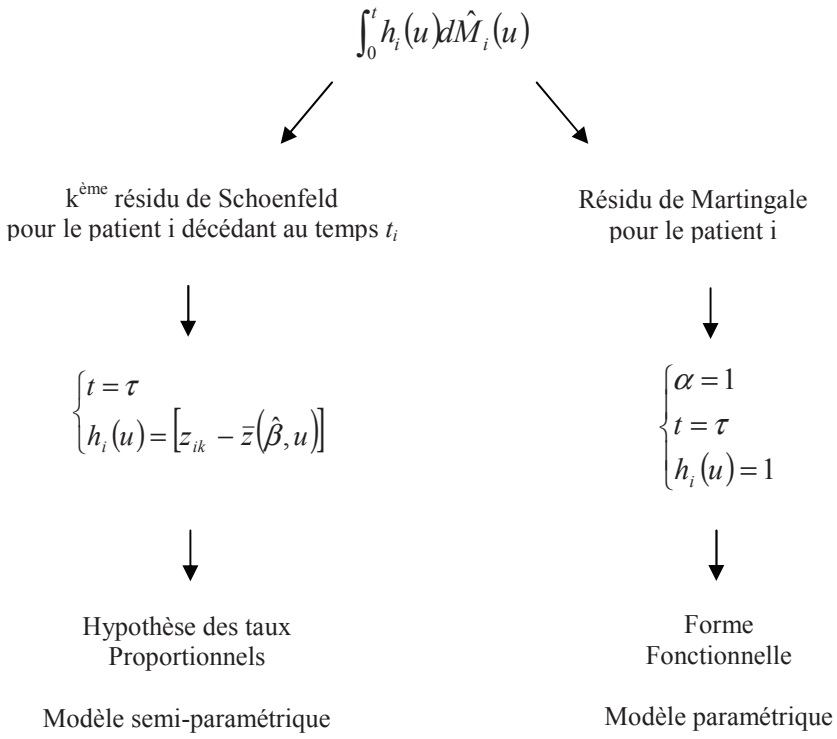


Figure III.13. Résidus reposant sur des processus dans le cadre de la survie globale

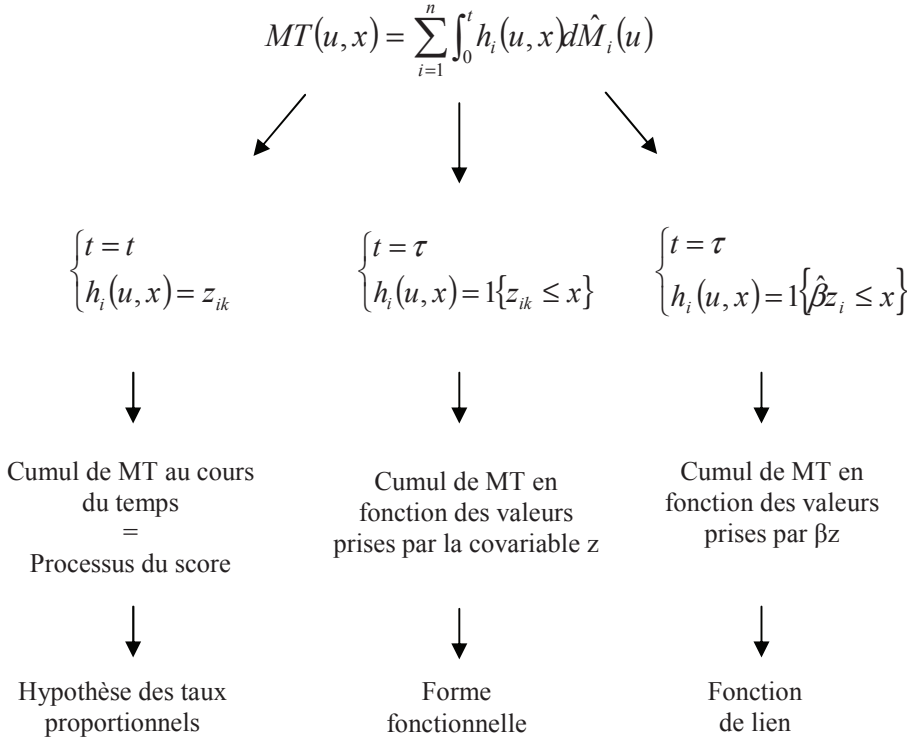


Figure III.14. Résidus calculés individuellement puis cumulés au cours du temps dans le cadre de la survie nette

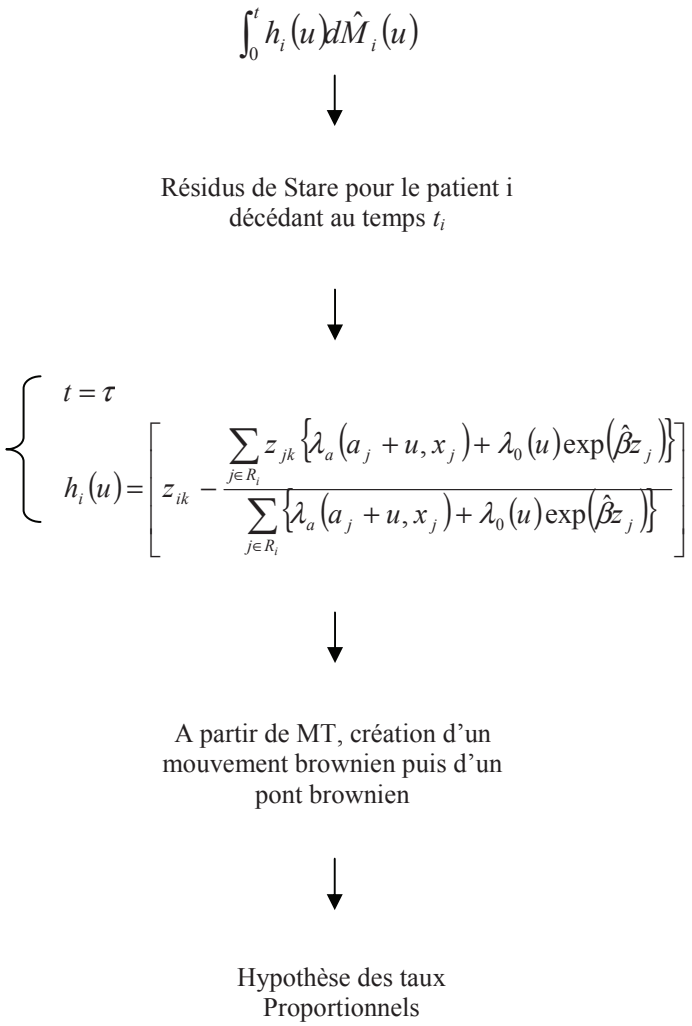
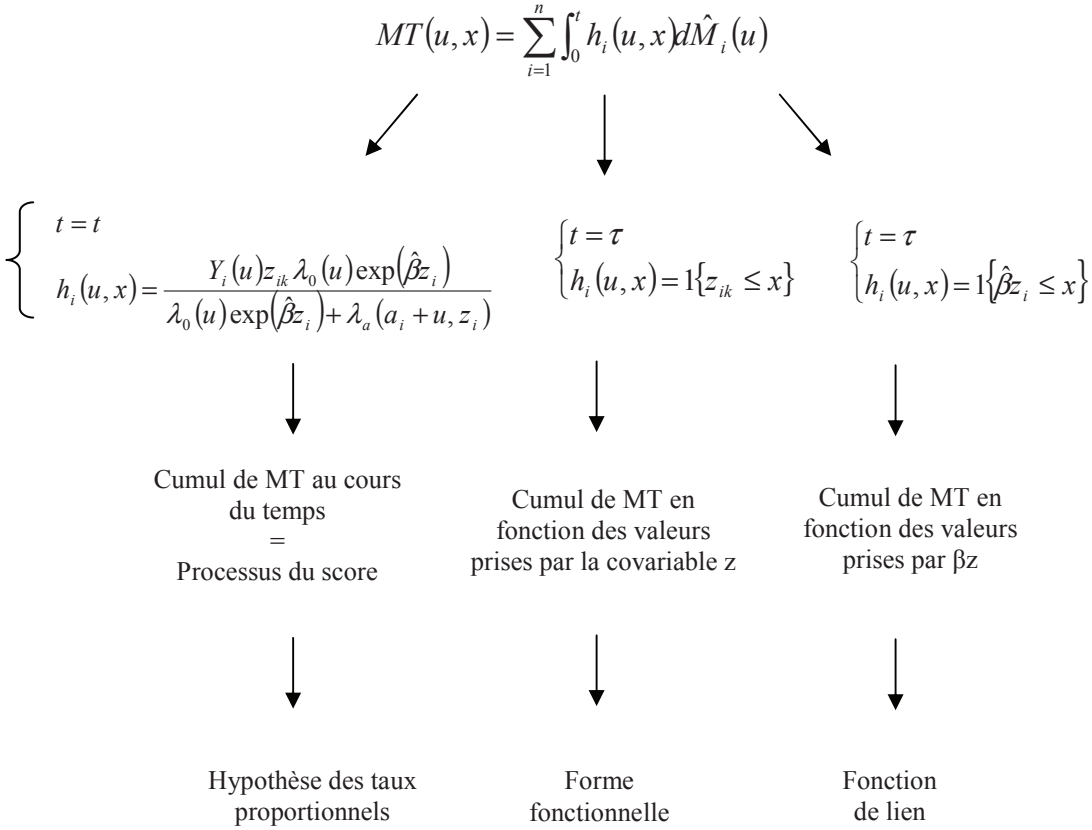


Figure III.15. Résidus reposant sur des processus dans le cadre de la survie nette



III.5. Performances des différents tests proposés

La performance des tests dans leur capacité à détecter la mauvaise spécification d'intérêt (effet non-proportionnel d'une covariable, forme fonctionnelle incorrecte ou fonction de lien incorrecte) seront évaluées dans le cadre de la survie nette, à l'aide de simulations. Faute de temps, nous présenterons ici uniquement les résultats concernant le test reposant sur le processus du score permettant de tester l'hypothèse de proportionnalité. Les résultats du test de Stare seront également présentés, ce qui permettra de situer le test que nous proposons par rapport au seul test actuellement disponible pour tester l'hypothèse de proportionnalité dans le cadre de la survie nette. L'effet de la covariable *âge*, l'une des covariables ayant le plus d'impact sur la survie, sera analysé (i) lorsque les autres hypothèses du modèle seront correctes et (ii) lorsque certaines d'entre elles seront incorrectes.

III.5.1. Design des simulations

Pour évaluer les performances des différentes méthodes, les quatre configurations que nous pouvons rencontrer en ce qui concerne l'effet d'une covariable continue sur le taux de mortalité en excès ont été simulées : Effet Linéaire et Proportionnel (LIN-PH), effet Linéaire et Non-Proportionnel (LIN-NPH), effet Non-Linéaire et Proportionnel (NLIN-PH), effet Non-Linéaire et Non-Proportionnel (NLIN-NPH). Dans chacun des cas, la covariable *sexe* suivait une loi uniforme qui générait autant d'hommes que de femmes, la covariable *année de diagnostic* suivant une loi uniforme entre 1980 et 1990 et concernant la covariable *âge*, 25% avaient un âge compris dans l'intervalle [30 ; 64], 35% avaient un âge compris dans l'intervalle [65 ; 74] et 40% avaient un âge compris dans l'intervalle [75 ; 85]. Cette distribution correspond approximativement à celle observée pour l'incidence du cancer du Côlon en France. Pour simuler l'effet LIN-PH et l'effet NLIN-PH de la covariable âge, le temps de survie dû au cancer, T_c, a été généré en utilisant la méthode de la transformée inverse [Ross, 2006]. En revanche, pour simuler un effet LIN-NPH et NLIN-NPH, T_c a été générer en utilisant le package PermAlgo [Sylvestre, 2008]. La simulation de chaque configuration est décrite en Table III.4 et Figures de l'Appendix 7. Le temps de survie dû aux autres causes, T_a, a été simulé à l'aide d'une loi exponentielle par morceaux obtenue à partir des tables de mortalité de la population générale. La date de point a été fixée en 1995 ; le

temps potentiel de suivi du patient *i* est donc égale à 1995 moins l'année de diagnostic de ce patient.

Le temps d'observation final T est égal au minimum entre $\{T_c, T_a, C\}$. Pour chaque scénario, 1000 jeux de données contenant 500 patients chacun ont été générés.

Table III.4. Paramètres de simulations

	T_c **Temps de survie dû au cancer**	β_{agec} **Effet de l'âge centré**	β_{agec^2} **Effet du carré de l'âge centré**
LIN-PH	Distribution Log-normal ($\mu = 0.875$, $\sigma = 1.37$)	0.05	-
NLIN-PH	Distribution Log-normal ($\mu = 0.875$, $\sigma = 1.37$)	0.05	0.0015
LIN-NPH	Distribution non spécifié	[(1/14).exp(-0.35.t)+0.01]	-
NLIN-NPH	Distribution non spécifié	[(1/14).exp(-0.35.t)+0.01]	0.0015

III.5.2. Evaluation de la performance des méthodes

Pour évaluer la performance des méthodes dans leur aptitude à détecter la mauvaise spécification de la covariable d'intérêt dans le modèle, nous estimerons le risque de première espèce α qui représente la probabilité de rejeter à tort l'hypothèse nulle alors qu'elle est vraie, et le risque de seconde espèce β qui représente la probabilité de ne pas rejeter l'hypothèse nulle alors qu'elle est fausse, et qui permet d'estimer la puissance du test (1 - β).

L'estimation de tous les critères de performances sera faite en ajustant le modèle de Remontet LIN-PH et le modèle de Remontet NLIN-PH [Remontet, 2007] sur les données simulées décrite dans la table III.4. Les scénarios simulant une proportionnalité de l'effet de l'âge nous permettront d'évaluer le risque de première espèce α et les scénarios simulant une non-proportionnalité de l'effet de l'âge nous permettront d'évaluer la puissance (1 - β). Ces deux critères seront évalués sur les 1000 jeux de données à l'aide de différents scénarios.

III.5.3. Résultats du test permettant de vérifier l'hypothèse des taux proportionnels

Les résultats sont présentés dans la table III.5. Les différents scénarios ainsi que le critère de performance associé qui va être évalué sont présentés en ligne et les résultats des deux méthodes considérées dans ce cas sont présentés en colonne. Les résultats obtenus avec le test de Stare permettront de situer le test que nous proposons par rapport au seul test actuellement disponible.

Table. III.5 : Analyse de performance du test reposant sur le processus du score pour étudier l'hypothèse des taux proportionnels

	Données simulées	Modèle ajusté	Critères de performance	Test proposé	Test de Stare
Scenario 1	LIN-PH	LIN-PH	α	4.4	5.0
Scenario 2	LIN-NPH	LIN-PH	1 - β	84.2	84.4
Scenario 3	NLIN-PH	NLIN-PH	α	5.1	4.7
Scenario 4	NLIN-NPH	NLIN-PH	1 - β	81.8	96.5
Scenario 5	NLIN-PH	LIN-PH	α	13.2	11.3
Scenario 6	NLIN-NPH	LIN-PH	1 - β	94.3	98.5

Nous pouvons observer que lorsque le modèle ajusté correspond à celui qui a généré les données (scenario 1 et scenario 3), le risque de première espèce est proche de 5% pour les deux méthodes.

Lorsque le modèle ajusté est différent du « vrai » modèle (caractère non-linéaire non spécifié dans le scénario 5), l'erreur de première espèce est supérieure à 5% pour les deux méthodes : 13.2% pour le test que nous proposons et 11.3% pour le test de Stare. Les tests semblent détecter la mauvaise spécification de l'effet non-linéaire de la covariable.

Dans le scénario 2, lorsque l'effet non-proportionnel n'est pas pris en compte, la puissance des deux tests est similaire ; elles sont proches de 84%. Cependant, nous pouvons voir que pour le scenario 4, les deux tests ne fournissent pas la même puissance alors que nous aurions pensé que ce scénario aurait confirmé les résultats obtenu pour le scénario 2.

Dans le scenario 6, les puissances sont supérieures à celles présentées précédemment ; elles sont également perturbées par la non prise en compte de la non-linéarité. Tout comme

pour le scénario 5, le fait d'ignorer la partie non-linéaire de l'effet nous amène à croire à un effet non-proportionnel ; c'est pourquoi la puissance augmente de façon artificielle.

III.5.4. Discussion

L'ensemble des résultats obtenus avec le test proposé nous semblent cohérents. Lorsque l'hypothèse H_0 est correcte (scénario 1 et 3), elle est rejetée à 5% et lorsque l'hypothèse H_0 n'est pas correcte, les autres hypothèses étant correctes (scénario 2 et 4), la puissance obtenue est supérieure à 80% ce qui est satisfaisant.

Les scénarios 5 et 6 montrent que lorsque d'autres hypothèses que la proportionnalité (non-linéarité ici) sont mal spécifiées, les tests peuvent être perturbés. Les valeurs importantes obtenues pour le risque de première espèce et pour la puissance sont dues au fait qu'ignorer la partie non-linéaire de l'effet nous amène à croire à un effet non-proportionnel; ce résultat de portée assez générale avait déjà été émis par Abrahamovicz pour d'autres tests [Abrahamovicz, 2007].

Un résultat concernant le test de Stare nous interpelle cependant : l'augmentation de la puissance que nous pouvons observer pour la méthode de Stare entre le scénario 2 et le scénario 4 est surprenante. Les raisons de cette augmentation sont en cours d'exploration et semblerait être liées à l'allure que prend la moyenne pondérée de la covariable âge au cours du temps. Il reste à confirmer que ce résultat, c'est-à-dire, $(1 - \beta)_{\text{Stare}} >> (1 - \beta)_{\text{Test proposé}}$, un peu surprenant, est retrouvé avec d'autres paramètres de simulation.

III.6. Application sur données réelles

Dans cette partie, nous illustrons la méthode proposée pour tester l'hypothèse des taux proportionnels sur quatre sites de cancer qui nous permettent de retrouver les scénarios étudiés dans l'étude de simulation. Les sites choisis sont les glandes salivaires, l'oropharynx, la maladie de hodgkin et le sein. Ils ont été choisis car des études précédentes ont montré que l'effet de l'âge était respectivement Linéaire et Proportionnel (LIN-PH), Linéaire et Non-Proportionnel (LIN-NPH), Non-Linéaire et Proportionnel (NLIN-PH) et Non-Linéaire et Non-Proportionnel (NLIN-NPH). Pour chacun de ces sites, tous les patients de plus de 15 ans diagnostiqués entre le 1er Janvier 1989 et le 31 Décembre 1997 par 16 registres français du réseau FRANCIM ont été inclus. La censure administrative était fixée au 1er Janvier 2002.

Le premier cas concerne les glandes salivaires (connu donc pour être LIN-PH), un modèle paramétrique du taux en excès linéaire et proportionnel a été ajusté sur les données, (référence au scénario 1). Le test de l'hypothèse des taux proportionnel, en considérant les autres hypothèses comme correctes, a fourni une p-value égale à 0.67 (la p-value reportée ici repose sur 500 processus gaussiens simulés). Ce résultat, illustré en figure 2a, suggère que la modélisation proportionnelle de l'effet de l'âge est adéquate aux données dans ce cas. Cela conforte les résultats obtenus dans les études précédentes.

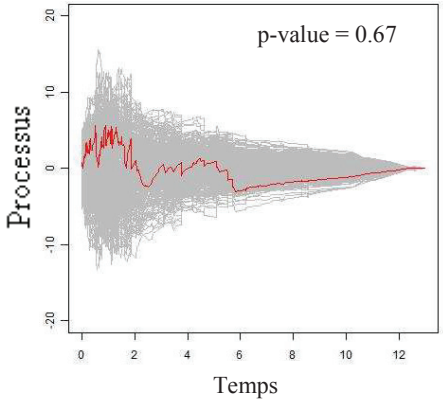
Concernant l'oropharynx (LIN-NPH), un modèle paramétrique du taux en excès linéaire et proportionnel a été ajusté sur les données (référence au scénario 2). Le test de l'hypothèse des taux proportionnel, en considérant les autres hypothèses comme correctes, a fourni une p-value très inférieure à 0.05 (sur les 500 processus gaussiens simulés, aucun maximum n'est supérieur au maximum du processus observé, la p-value est donc nulle). Ce résultat, illustré en figure 2b, suggère que la modélisation proportionnelle de l'effet de l'âge est inadéquate aux données dans ce cas. Graphiquement, nous pouvons voir que le processus du score est en dehors du « nuage » de processus gaussiens simulés au cours des deux premières années qui suivent le diagnostic. Notons que dans ce cas particulier, le maximum du processus du score observé étant très grand, nous pouvons en déduire directement les résultats du test (H_0 rejetée). Cela conforte les résultats obtenus dans les études précédentes.

Concernant la maladie de Hodgkin (NLIN-PH), un modèle paramétrique du taux en excès linéaire et proportionnel a été ajusté sur les données (référence au scénario 5). Le test de l'hypothèse des taux proportionnel, en considérant les autres hypothèses comme correctes, a fourni une p-value très inférieure à 0.05 (sur les 500 processus gaussiens simulés, aucun

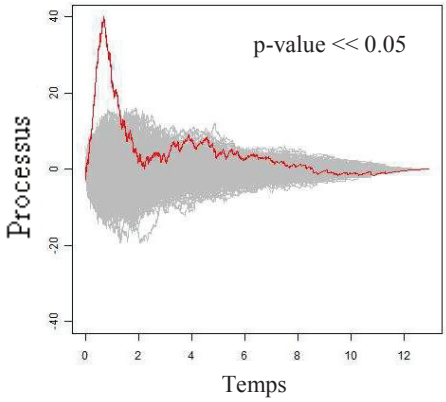
maximum n'est supérieur au maximum du processus observé, la p-value est donc nulle). Ce résultat, illustré en figure 2c, suggère que la modélisation proportionnelle de l'effet de l'âge est inadéquate aux données dans ce cas. Graphiquement, nous pouvons voir que le processus du score est en dehors du « nuage » de processus gaussiens simulés au cours des deux premières années qui suivent le diagnostic. Ce résultat va à l'encontre des résultats obtenus dans les études précédentes. Le fait que le caractère non-linéaire ne soit pas pris en compte amène à rejeter l'hypothèse de proportionnalité. Cet exemple fait référence au scénario 5 de l'étude de simulation.

Enfin, le dernier cas concerne le sein (NLIN-NPH). Un modèle paramétrique du taux en excès linéaire et proportionnel a été ajusté sur les données. Le test de l'hypothèse des taux proportionnel, en considérant les autres hypothèses comme correctes, a fourni une p-value égale à 0.002 (figure 2d). Ce test fournit un résultat que l'on attendait mais chacun doit être attentif au fait qu'ignorer le caractère non-linéaire « parasite » les performances du test.

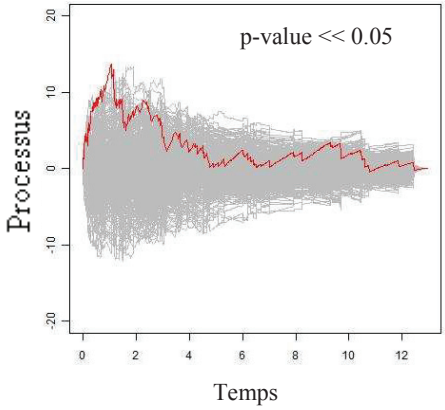
Figure III.16. Graphe des processus du score au cours du temps pour les quatre sites de cancers étudiés (le score observé est en rouge, les processus gaussiens simulés en gris, la p-value a été calculée sur 500 processus gaussiens)



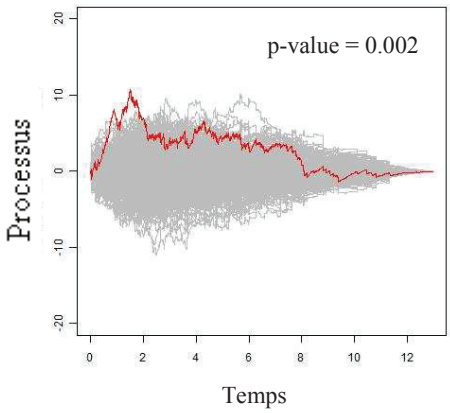
(a) Glandes salivaires



(b) Oropharynx



(d) - Hodgkin



(d) - Sein

III.7. Discussion et Conclusion

Dans cette partie, nous proposons une boîte à outils permettant de vérifier les différentes hypothèses d'un modèle paramétrique du taux en excès. Les tests appartenant à cette boîte à outils sont des tests formels qui ont été développés dans le même cadre théorique.

Les premiers résultats obtenus pour le test permettant de vérifier l'hypothèse de proportionnalité sont encourageants. L'erreur de première espèce est correcte et la puissance est proche de celle observée pour le test de Stare.

D'un point de vue pratique, lors de l'utilisation de ce test sur données réelles, les résultats doivent être interprétés avec précaution du fait que la mauvaise spécification d'un aspect du modèle peut avoir un impact sur le résidu permettant de décrire un autre aspect. Par exemple, la mauvaise spécification du caractère non linéaire de l'effet de la covariable d'intérêt peut impacter le comportement du processus du score servant à vérifier l'hypothèse de proportionnalité. Dans ce cas, comme il l'a été recommandé [Abrahamovicz, 2007], le test de l'hypothèse des taux proportionnel doit être effectué en spécifiant de manière systématique un effet non-linéaire pour la covariable d'intérêt.

Une solution permettant de remédier à ce problème serait de développer un outil permettant de tester de façon *simultanée* les différentes hypothèses du modèle, notamment les plus importantes, à savoir la proportionnalité et la forme fonctionnelle. Différents outils [Sasieni, 2003][Pohar, 2008] ont été développés dans le cadre de la survie globale. Une perspective intéressante mais complexe pourrait être d'adapter ces outils au cadre de la survie nette.

Ces premiers résultats obtenus ont permis également de vérifier que le rationnel et les programmes étaient correctes. Cependant, même si l'étude de simulation ne semble pas montrer d'incohérence, une étape importante pourrait être de valider, de façon plus théorique, les développements qui ont été fait.

Les résultats des tests portant sur la forme fonctionnelle et sur la fonction de lien ne sont pas encore disponibles, l'implémentation étant en cours. L'analyse de performance de ces deux tests sera faite selon les mêmes modalités d'évaluation que le test de vérification de l'hypothèse des taux proportionnels, une fois la phase d'implémentation terminée.

Pour chacun des tests, si l'hypothèse nulle est rejetée, les graphes des processus respectifs ne nous permettent pas d'obtenir d'information précise sur la forme à modéliser concernant le caractère non-proportionnel, non-linéaire ou la fonction de lien. Ils permettent

toutefois de voir respectivement pour quelle période, pour quelle valeur de la covariable ou pour quelle valeur du logarithme des taux relatifs, le modèle est mal spécifié.

Enfin, lorsque ces outils auront été implémentés et testés (tous ou partiellement), il sera nécessaire de les rendre disponibles en routine (en les intégrant à un package R), afin que l'étape de vérification des hypothèses du modèle, qui sera alors rapide, soit rendue accessible à tous. Dans ce but, des contacts ont déjà été pris avec M Pohar-Perme (responsable du package *relsurv*) et G Hédelin (responsable du package *Flexrsurv* développé dans la cadre du projet de recherche CENSUR) afin d'intégrer à terme nos outils dans leur package.

Appendix 1 : Expression du score partiel issu du modèle de Cox à l'aide de transformées de Martingale

Soit le modèle de Cox suivant :

$$\lambda(t,z) = \lambda_0(t)\exp(\beta z)$$

avec λ le taux de mortalité observé, λ_0 le taux de mortalité de base, z le vecteur des covariables incluses dans le modèle et β le vecteur coefficient des covariables.

En utilisant les processus de comptage, la vraisemblance partielle du modèle de Cox s'exprime de la manière suivante :

$$V^*(\beta) = \prod_{i=1}^{n}\prod_{t\geq 0}\left[\frac{Y_i(t)\exp(\beta z_i)}{\sum_j Y_j(t)\exp(\beta z_j)}\right]^{dN_i(t)}$$

La $k^{\text{ième}}$ composante du vecteur score issu du modèle de Cox est alors :

$$U_k(\beta,\tau) = \frac{d\log(V^*(\beta))}{d\beta_k} = \sum_i \int_0^\tau \left[z_{ik} - \frac{\sum_j Y_j(u) z_{jk}\exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)}\right] dN_i(u)$$

Le processus du score associé à la $k^{\text{ième}}$ covariable est alors :

$$U_k(\beta,t) = \sum_i \int_0^t \left[z_{ik} - \frac{\sum_j Y_j(u) z_{jk}\exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)}\right] dN_i(u)$$

Or,

$$U_k(\beta,t)=\sum_i\int_0^t\left[z_{ik}-\frac{\sum_j Y_j(u)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)}\right]dN_i(u)$$

$$U_k(\beta,t)=\sum_i\int_0^t\left[z_{ik}-\frac{\sum_j Y_j(u)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)}\right][dN_i(u)-Y_i(u)\lambda_i(u)du]+\sum_i\int_0^t\left[z_{ik}-\frac{\sum_j Y_j(u)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)}\right]Y_i(u)\lambda_i(u)du$$

$$U_k(\beta,t)=\sum_i\int_0^t\left[z_{ik}-\frac{\sum_j Y_j(u)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)}\right]dM_i(u)+\sum_i\int_0^t\left[z_{ik}-\frac{\sum_j Y_j(u)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)}\right]Y_i(u)\lambda_i(u)du$$

Et,

$$\sum_i\int_0^t\left[z_{ik}-\frac{\sum_j Y_j(u)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)}\right]Y_i(u)\lambda_i(u)du=\sum_i\int_0^t z_{ik}Y_i(u)\lambda_i(u)du-\sum_i\int_0^t\left[\frac{\sum_j Y_j(u)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)}\right]Y_i(u)\lambda_i(u)du$$

$$\sum_i\int_0^t\left[z_{ik}-\frac{\sum_j Y_j(u)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)}\right]Y_i(u)\lambda_i(u)du=\sum_i\int_0^t z_{ik}Y_i(u)\lambda_i(u)du-\int_0^t\left[\frac{\sum_j Y_j(u)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)}\right]\sum_i Y_i(u)\lambda_i(u)du$$

$$\sum_i\int_0^t\left[z_{ik}-\frac{\sum_j Y_j(u)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)}\right]Y_i(u)\lambda_i(u)du=\sum_i\int_0^t z_{ik}Y_i(u)\lambda_i(u)du-\int_0^t\left[\frac{\sum_j Y_j(u)z_{jk}\lambda_0(u)\exp(\beta z_j)}{\sum_j Y_j(u)\lambda_0(u)\exp(\beta z_j)}\right]\sum_i Y_i(u)\lambda_i(u)du$$

$$\sum_i\int_0^t\left[z_{ik}-\frac{\sum_j Y_j(u)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)}\right]Y_i(u)\lambda_i(u)du=0$$

Cela démontre que le processus du score pour la covariable *k* peut s'écrire comme une transformée de martingale de la façon suivante :

$$U_k(\beta,t)=\sum_i\int_0^t\left[z_{ik}-\frac{\sum_j Y_j(u)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)}\right]dM_i(u)$$

avec $h_i(u,x)=\left[z_{ik}-\frac{\sum_j Y_j(u)z_{jk}\exp(\beta z_j)}{\sum_j Y_j(u)\exp(\beta z_j)}\right]$

Appendix 2 : Distribution du processus du score pour un modèle paramétrique

Soit le modèle paramétrique suivant :

$$\lambda(t,z) = \lambda_0(t)\exp(\beta z)$$

avec λ le taux de mortalité observé, λ_0 le taux de mortalité de base, z le vecteur des covariables incluses dans le modèle et β le vecteur coefficient des covariables.

2.1. Processus du score sous un modèle paramétrique

Pour un modèle paramétrique de la forme précédente, la log-vraisemblance s'écrit de la manière suivante :

$$\begin{aligned}
L(\beta,t) &= \sum_i \log\left(S_i(t)\lambda_i(t)^{\delta_i}\right) \\
L(\beta,t) &= \sum_i \left[\log(S_i(t)) + \delta_i \log(\lambda_i(t))\right] \\
L(\beta,t) &= \sum_i \left[-\Lambda_i(t) + \delta_i \log(\lambda_i(t))\right] \\
L(\beta,t) &= \sum_i L_i(\beta,t)
\end{aligned}$$

La dérivée de la log-vraisemblance au temps *t*, qui représente le processus du score au temps *t*, s'écrit de la manière suivante :

$$U_k(\beta,t)=\sum_i U_{ik}(\beta,t)$$
$$U(\beta,t)=\frac{\partial L(\beta,t)}{\partial \beta_k}$$
$$U(\beta,t)=\sum_i \frac{\partial L_i(\beta,t)}{\partial \beta_k}$$
$$U(\beta,t)=\sum_i \left[-z_{ik}\Lambda_i(t)+\delta_i \frac{z_{ik}\lambda_i(t)}{\lambda_i(t)} \right]$$
$$U(\beta,t)=\sum_i \left[-z_{ik}\int_0^t \lambda_i(u)du + z_{ik}\int_0^t dN_i(u) \right]$$
$$U(\beta,t)=\sum_i \left[z_{ik}\left[\int_0^t dN_i(u) - \int_0^t \lambda_i(u)du \right] \right]$$
$$U(\beta,t)=\sum_i \left[\int_0^t z_{ik}\, dM_i(u) \right]$$

Nous reconnaissons alors ici une transformée de martingale avec la fonction prévisible associée $h_i(t)=z_{ik}$.

2.2. Approximation de la distribution du processus du score sous l'hypothèse nulle

En utilisant le développement de Taylor pour la variable β,

$$U(\hat{\beta},t)\approx U(\beta_0,t)+(\hat{\beta}-\beta_0)I(\hat{\beta},t)$$
$$n^{-1/2}U(\hat{\beta};t)\approx n^{-1/2}U(\beta_0;t)-n^{-1}I(\hat{\beta};t)n^{1/2}(\hat{\beta}-\beta_0) \qquad \text{(III.9)}$$

Or, si le modèle est correct, le processus du score au temps τ = fin de suivi est nul :

$$n^{-1/2}U(\hat{\beta};\tau) \approx n^{-1/2}U(\beta_0;\tau) - n^{-1}I(\hat{\beta};\tau)n^{1/2}(\hat{\beta}-\beta_0)$$
$$0 \approx n^{-1/2}U(\beta_0;\tau) - n^{-1}I(\hat{\beta};\tau)n^{1/2}(\hat{\beta}-\beta_0)$$
$$n^{1/2}(\hat{\beta}-\beta_0) \approx n^{-1/2}U(\beta_0;\tau)\times nI(\hat{\beta};\tau)^{-1}$$

Donc, en remplaçant $n^{1/2}(\hat{\beta}-\beta)$ dans l'équation (III.9) par $n^{-1/2}U(\beta_0;\tau)\times nI(\hat{\beta};\tau)^{-1}$, on a :

$$n^{-1/2}U(\hat{\beta};t) \approx n^{-1/2}U(\beta_0;t) - n^{-1}I(\hat{\beta};t)n^{-1/2}U(\beta_0;\tau)\times nI(\hat{\beta};\tau)^{-1}$$
$$n^{-1/2}U(\hat{\beta};t) \approx n^{-1/2}\left(U(\beta_0;t) - I(\hat{\beta};t)I(\hat{\beta};\tau)^{-1}U(\beta_0;\tau)\right)$$

La distribution de $n^{-1/2}U(\hat{\beta},t)$ est donc asymptotiquement équivalente à la distribution du processus W_z, tel que :

$$W_z(t) = n^{-1/2}\left(U(\beta_0;t) - I(\hat{\beta};t)I(\hat{\beta};\tau)^{-1}U(\beta_0;\tau)\right)$$

Cependant, nous connaissons la distribution du processus du score observé $U(\hat{\beta},t)$ mais pas celle du processus du score sous l'hypothèse nulle $U(\beta_0,t)$.

En effet, sous l'hypothèse nulle

$$U_k(\beta_0,t) = \sum_i \left[\int_0^t z_{ik}\,dM_i(u)\right]$$
$$U_k(\beta_0,t) = \sum_i \left[\int_0^t z_{ik}\,d\left(N_i(u) - Y_i(u)\lambda_0(u)\exp(\beta_0 z_i)du\right)\right]$$

Or, M_i étant fonction de λ_0 et de β_0 et ces derniers n'étant pas connu, M_i n'est donc pas non plus connu sous l'hypothèse nulle. Cependant, sa distribution peut être approximée par le processus gaussien $N_i(u).G_i$ avec G_i suivant une loi normale standardisée. Ce processus possède les mêmes caractéristiques que M_i concernant son espérance et sa variance [Fleming

& Harrington, 1991][Lin, 1993]. $U_k(\beta_0;t)$ pourra alors être approximé par $\hat{M}_{1k}(t)$ tel que pour le modèle de Cox, $\hat{M}_{1k}(t)$ est égale à :

$$\hat{M}_{1k}(t) = \sum_i \int_0^t \left[z_{ik} - \frac{\sum_j Y_j(u) z_{jk} \exp\left(\hat{\beta} z_j\right)}{\sum_j Y_j(u) \exp\left(\hat{\beta} z_j\right)} \right] dN_i(u) G_i$$

et pour le modèle paramétrique de type (III.1), $\hat{M}_{1k}(t)$ est égale à :

$$\hat{M}_{1k}(t) = \sum_i \int_0^t z_{ik} dN_i(u) G_i$$

Le processus du score peut donc être approximé par :

$$\tilde{W}_z(t,z) = n^{-1/2}\left(\hat{M}_1(t) - I\left(\hat{\beta};t\right) I\left(\hat{\beta};\tau\right)^{-1} \hat{M}_1(\tau) \right)$$

2.3. Equivalence des processus $W_z(t,z)$ et $\tilde{W}_z(t,z)$

L'expression du processus W_z est la suivante :

$$W_z(t) = n^{-1/2}\left(U(\beta_0;t) - I\left(\hat{\beta};t\right) I\left(\hat{\beta};\tau\right)^{-1} U(\beta_0;\tau) \right)$$

Il peut alors s'exprimer tel que :

$$W_z(t) = n^{-1/2}\left(\sum_i \int_0^t \frac{\partial \log \lambda_{obs}(u)}{\partial \beta} dM_i(u) \Big|_{\beta=\beta_0} - I\left(\hat{\beta};t\right) I\left(\hat{\beta};\tau\right)^{-1} \sum_i \int_0^\tau \frac{\partial \log \lambda_{obs}(u)}{\partial \beta} dM_i(u) \Big|_{\beta=\beta_0} \right)$$

$$W_z(t) = n^{-1/2} \sum_i \int_0^\tau h_i(\beta_0, t, u, z) dM_i(u)$$

avec

$$h_i(\beta_0,t,u,z) = 1(u \le t) \times \left.\frac{\partial \log \lambda_{obs}(u)}{\partial \beta}\right|_{\beta=\beta_0} - I(\hat{\beta};t) I(\hat{\beta};\tau)^{-1} \left.\frac{\partial \log \lambda_{obs}(u)}{\partial \beta}\right|_{\beta=\beta_0}$$

Or M(t) étant une martingale telle que $E(M(t)^2) < +\infty$ pour tout $t \geq 0$, il existe un unique processus prévisible croissant, cadlag, appelé $\langle M(t) \rangle$, tel que $M(t)^2 - \langle M(t) \rangle$ soit une martingale.

$\langle M(t) \rangle$ est un processus à variations prévisibles tel que $E(\langle M(t) \rangle) < +\infty$ et $\langle M(0) \rangle = 0$; il est le compensateur de M^2.

$$E\left(M^2(t) - \langle M(t) \rangle\right) = E\left(M_0^2(t) - \langle M(0) \rangle\right)$$
$$E\left(M^2(t) - \langle M(t) \rangle\right) = 0$$

Alors,

$$E\left(M^2(t)\right) = E\left(\langle M(t) \rangle\right)$$

Le processus W_z étant une transformée de martingale, il a les propriétés d'une martingale et est une martingale ; conditionnellement aux covariables, nous pouvons donc écrire :

$$E\left(\left[n^{-1/2} \sum_i \int_0^\tau h_i(\beta_0,t,u,z) dM_i(u)\right]^2\right) = E\left(\left\langle n^{-1/2} \sum_i \int_0^\tau h_i(\beta_0,t,u,z) dM_i(u) \right\rangle\right)$$
$$E\left(\left[n^{-1/2} \sum_i \int_0^\tau h_i(\beta_0,t,u,z) dM_i(u)\right]^2\right) = E\left(n^{-1} \sum_i \int_0^\tau h_i^2(\beta_0,t,u,z) d\langle M_i(u) \rangle\right)$$
$$E\left(\left[n^{-1/2} \sum_i \int_0^\tau h_i(\beta_0,t,u,z) dM_i(u)\right]^2\right) = E\left(n^{-1} \sum_i \int_0^\tau h_i^2(\beta_0,t,u,z) Y_i(u) \lambda_{obs}(u) du\right)$$

Or,

$$Var\left(\left[n^{-1/2}\sum_i\int_0^\tau h_i(\beta_0,t,u,z)dM_i(u)\right]\right)=E\left(\left[n^{-1/2}\sum_i\int_0^\tau h_i(\beta_0,t,u,z)dM_i(u)\right]^2\right)$$

En développant,

$$Var\left(\left[n^{-1/2}\sum_i\int_0^\tau h_i(\beta_0,t,u,z)dM_i(u)\right]\right)=E\left(n^{-1}\sum_i\int_0^\tau h_i^2(\beta_0,t,u,z)Y_i(u)\lambda_{obs}(u)du\right)$$

$$Var\left(\left[n^{-1/2}\sum_i\int_0^\tau h_i(\beta_0,t,u,z)dM_i(u)\right]\right)=E\left(\int_0^\tau h_i^2(\beta_0,t,u,z)Y_i(u)\lambda_{obs}(u)du\right)$$

En ce qui concerne le processus $\widetilde{W}(t)$,

$$\widetilde{W}_z(t)\approx n^{-1/2}\sum_i\int_0^\tau h_i(\beta_0,t,u,z)dN_i(u)G_i$$

$$Var\left(n^{-1/2}\sum_i\int_0^\tau h_i(\beta_0,t,u,z)dN_i(u)G_i\right)=n^{-1}Var\left(\sum_i\int_0^\tau h_i(\beta_0,t,u,z)dN_i(u)G_i\right)$$

$$Var\left(n^{-1/2}\sum_i\int_0^\tau h_i(\beta_0,t,u,z)dN_i(u)G_i\right)=n^{-1}\sum_i Var\left(\int_0^\tau h_i(\beta_0,t,u,z)dN_i(u)G_i\right)$$

Covariables iid

Or,

$$Var\left(\int_0^\tau h_i(\beta_0,t,u,z)dN_i(u)G_i\right)=Var\left(\int_0^\tau h_i(\beta_0,t,u,z)dN_i(u)\right)+\left[E\left(\int_0^\tau h_i(\beta_0,t,u,z)dN_i(u)\right)\right]^2$$

Alors,

$$Var\left(n^{-1/2}\sum_i \int_0^\tau h_i(\beta_0,t,u,z)dN_i(u)G_i \right) = n^{-1}\sum_i \left[Var\left(\int_0^\tau h_i(\beta_0,t,u,z)dN_i(u) \right) + \left[E\left(\int_0^\tau h_i(\beta_0,t,u,z)dN_i(u) \right) \right]^2 \right]$$

$$Var\left(n^{-1/2}\sum_i \int_0^\tau h_i(\beta_0,t,u,z)dN_i(u)G_i \right) = n^{-1}\sum_i \left[E\left(\left(\int_0^\tau h_i(\beta_0,t,u,z)dN_i(u) \right)^2 \right) \right]$$

$$Var\left(n^{-1/2}\sum_i \int_0^\tau h_i(\beta_0,t,u,z)dN_i(u)G_i \right) = n^{-1}\sum_i \left[E\left(\int_0^\tau h_i^2(\beta_0,t,u,z)dN_i(u) \right) \right]$$

Conditionnellement aux covariables, en utilisant la loi des grands nombres et puisque $E(N_i) = E(\Lambda_i)$:

$$Var\left(n^{-1/2}\sum_i \int_0^\tau h_i(\beta_0,t,u,z)dN_i(u)G_i \right) \underset{n\to+\infty}{\to} E\left(\int_0^\tau h_i^2(\beta_0,t,u,z)Y_i(u)\lambda_i(u)du \right)$$

La variance du processus $\tilde{W}(t)$ converge donc asymptotiquement vers la variance de $W(t)$.

Appendix 3 : Résidus de martingale cumulés pour la vérification de la forme fonctionnelle de la covariable d'intérêt

Pour alléger l'écriture, nous supposerons qu'il n'y a qu'une seule covariable incluse dans le modèle.

Soit M_{cum} le processus des résidus de martingale cumulés pour la covariable incluse dans le modèle :

$$M_{cum}(x,\beta) = \sum_{i=1}^{n} \int_0^{\tau} I(z_i \le x) dM_i(t)$$

avec $M_i(t) = N_i(t) - \int_0^t Y_i(u)\lambda(u)du$

En utilisant le développement de Taylor sur cette formule, on obtient le développement suivant :

$$n^{1/2} M_{cum}\left(x,\hat{\beta}\right) \approx n^{1/2} M_{cum}(x,\beta_0) + n^{1/2}\left(\hat{\beta}-\beta_0\right) \sum_{i=1}^{n} \int_0^{\tau} I(z_i \le x) \frac{\partial d\hat{M}_i(t)}{\partial \beta_0}$$

En utilisant l'approximation faite sur le processus du score,

$$n^{1/2}\left(\hat{\beta}-\beta_0\right) \approx n^{-1/2} U(\beta_0;\tau) \times nI\left(\hat{\beta};\tau\right)^{-1}$$

et en développant le terme possédant une intégrale,

$$\frac{\partial d\hat{M}_i(t)}{\partial \beta_0} = \frac{\partial dN_i(t)}{\partial \beta_0} - \frac{\partial d\Lambda_i(t,z_i,\beta_0)}{\partial \beta_0}$$
$$\frac{\partial d\hat{M}_i(t)}{\partial \beta_0} = -\frac{\partial d\Lambda_i(t,z_i,\beta_0)}{\partial \beta_0}$$

avec

$$-\frac{\partial d\Lambda_i(t,z_i,\beta_0)}{\partial \beta_0} = -\frac{\partial \exp(\beta_0 z_i) d\Lambda_0(t)}{\partial \beta_0}$$
$$-\frac{\partial d\Lambda_i(t,z_i,\beta_0)}{\partial \beta_0} = -z_i \exp(\beta_0 z_i) d\Lambda_0(t)$$

Nous obtenons,

$$n^{1/2} M_{cum}\left(x, \hat{\beta}\right) \approx n^{1/2} M_{cum}(x, \beta_0) - n^{-1/2} U(\beta_0; \tau) \times nI\left(\hat{\beta}; \tau\right)^{-1} \sum_{i=1}^{n} \int_0^{\tau} I(z_i \leq x) z_i \exp(\beta_0 z_i) d\Lambda_0(t)$$

$$n^{1/2} M_{cum}\left(x, \hat{\beta}\right) \approx n^{1/2} M_{cum}(x, \beta_0) - n^{-1/2} U(\beta_0; \tau) \times nI\left(\hat{\beta}; \tau\right)^{-1} \times J(\tau, z, \beta_0)$$

La distribution de $n^{1/2} M_{cum}\left(x, \hat{\beta}\right)$ est donc asymptotiquement équivalente à la distribution du processus W_z, tel que:

$$W_z(x) = n^{1/2} M_{cum}(x, \beta_0) - n^{-1/2} U(\beta_0; \tau) \times nI\left(\hat{\beta}; \tau\right)^{-1} \times J(\tau, z, \beta_0)$$

Appendix 4 : Résidus de martingale cumulés pour la vérification de la fonction de lien

Les résidus de martingales cumulées en fonction des valeurs du log du taux relatif βz s'expriment de la manière suivante :

$$M_{cum}(x,\beta) = \sum_{i=1}^{n} \int_0^{\tau} I(\beta z_i \le x) dM_i(u)$$

avec $M_i(t) = N_i(t) - \int_0^t Y_i(u)\lambda(u)du$

En utilisant le développement de Taylor sur la formule ci-dessus, on obtient le développement suivant :

$$n^{1/2} M_{cum}(x,\hat{\beta}) \approx n^{1/2} M_{cum}(x,\beta_0) + n^{1/2}(\hat{\beta} - \beta_0) \sum_{i=1}^{n} \int_0^{\tau} I(\beta z_i \le x) \frac{\partial d\hat{M}_i(t)}{\partial \beta_0}$$

En utilisant toujours l'approximation suivante :

$$n^{1/2}(\hat{\beta} - \beta_0) \approx n^{-1/2} U(\beta_0;\tau) \times nI(\hat{\beta};\tau)^{-1}$$

et en développant le terme possédant une intégrale et en posant $\gamma = \beta z_i$,

$$\frac{\partial d\hat{M}_i(t)}{\partial \beta_0} = \frac{\partial dN_i(t)}{\partial \beta_0} - \frac{\partial d\Lambda_i(t,\gamma_i,\beta_0)}{\partial \beta_0}$$
$$\frac{\partial d\hat{M}_i(t)}{\partial \beta_0} = -\frac{\partial d\Lambda_i(t,\gamma_i,\beta_0)}{\partial \beta_0}$$

avec

$$-\frac{\partial d\Lambda_i(t,\gamma_i,\beta_0)}{\partial \beta_0} = -\frac{\partial \exp(\beta_0\gamma_i) d\Lambda_0(t)}{\partial \beta_0}$$
$$-\frac{\partial d\Lambda_i(t,\gamma_i,\beta_0)}{\partial \beta_0} = -\gamma_i \exp(\beta_0\gamma_i) d\Lambda_0(t)$$

Nous obtenons,

$$n^{1/2} M_{cum}\left(x,\hat{\beta}\right) \approx n^{1/2} M_{cum}(x,\beta_0) - n^{-1/2} U(\beta_0;\tau) \times nI\left(\hat{\beta};\tau\right)^{-1} \sum_{i=1}^{n} \int_0^{\tau} I(\gamma_i \le x)\gamma_i \exp(\beta_0\gamma_i) d\Lambda_0(t)$$
$$n^{1/2} M_{cum}\left(x,\hat{\beta}\right) \approx n^{1/2} M_{cum}(x,\beta_0) - n^{-1/2} U(\beta_0;\tau) \times nI\left(\hat{\beta};\tau\right)^{-1} \times J(\tau,\gamma,\beta_0)$$

La distribution de $n^{1/2} M_{cum,k}\left(x,\hat{\beta}\right)$ est donc asymptotiquement équivalente à la distribution du processus W_z, tel que:

$$W_z(x) \approx n^{1/2} M_{cum,k}(x,\beta_0) - n^{-1/2} U(\beta_0;\tau) \times nI\left(\hat{\beta};\tau\right)^{-1} \times J(\tau,\gamma,\beta_0)$$

Appendix 5 : Processus du score pour un modèle paramétrique du taux en excès

Pour alléger l'écriture, nous supposerons qu'il n'y a qu'une seule covariable incluse dans le modèle.

Soit le modèle paramétrique du taux en excès suivant :

$$\lambda_{obs}(t,z) = \lambda_c(t,z) + \lambda_a(a+t,x)$$

λ_{obs} représentant le taux de mortalité observé au temps t, λ_c représentant le taux de mortalité en excès au temps t, λ_a représentant le taux de mortalité attendu au temps t, x représentant le vecteur des covariables démographiques et z représentant le vecteur des variables pronostiques incluses dans le modèle.

Par définition, la log-vraisemblance d'un modèle paramétrique de taux en excès est la suivante :

$$L_i(\beta) = -\int_0^{t_i} \lambda_c(u,z_i)du - \int_0^{t_i} \lambda_a(a_i+u,z_i)du + \delta_i \times \log[\lambda_c(t_i,z_i) + \lambda_a(a_i+t_i,x_i)]$$

Le score s'écrit alors de la manière suivante :

$$U(\beta) = \frac{\partial}{\partial\beta} L(\beta)$$

$$U(\beta) = \frac{\partial}{\partial\beta} \sum_i L_i(\beta)$$

$$U(\beta) = \sum_i \left[-\int_0^{t_i} \frac{\partial}{\partial\beta}(\lambda_c(u,z_i))du + \delta_i \times \frac{\partial}{\partial\beta}[\log[\lambda_c(t_i,z_i) + \lambda_a(a_i+t_i,x_i)]] \right]$$

$$U(\beta) = \sum_i \left[-\int_0^{t_i} z_i \lambda_0(u)\exp(\beta z_i)du + \delta_i \times \frac{z_i \lambda_0(t_i)\exp(\beta z_i)}{\lambda_0(t_i)\exp(\beta z_i) + \lambda_a(a_i+t_i,x_i)} \right]$$

$$U(\beta) = \sum_i \left[-\int_0^{t_i} z_i \lambda_0(u)\exp(\beta z_i)du + \delta_i \times \frac{z_i \lambda_0(t_i)\exp(\beta z_i)}{\lambda_0(t_i)\exp(\beta z_i) + \lambda_a(a_i+t_i,x_i)} \right]$$

En utilisant les processus de comptage et en posant τ la fin du suivi, le processus du score peut s'écrire alors de la manière suivante :

$$U(\beta)=\sum_i\left[-\int_0^\tau Y_i(u)z_i\lambda_0(u)\exp(\beta z_i)du+N_i(t_i)\times\frac{z_i\lambda_0(t_i)\exp(\beta z_i)}{\lambda_0(t_i)\exp(\beta z_i)+\lambda_a(a_i+t_i,x_i)}\right]$$

En posant la fonction $\alpha_i : s\mapsto\alpha_i(s)=\frac{z_i\lambda_0(s)\exp(\beta z_i)}{\lambda_0(s)\exp(\beta z_i)+\lambda_a(a_i+s,x_i)}$, on a :

$$U(\beta)=\sum_i\left[-\int_0^\tau Y_i(u)z_i\lambda_0(u)\exp(\beta z_i)du+N_i(t_i)\times\alpha(t_i)\right]$$

Or, les fonctions N_i et α_i étant à variations bornées, on peut utiliser la formule d'intégration par parties :

$$N_i(t_i)\alpha_i(t_i)-N_i(t_0)\alpha_i(t_0)=\int_0^{t_i}N_i(s^-)d\alpha_i(s)+\int_0^{t_i}\alpha_i(s)dN_i(s)$$

Les termes $N_i(t_0)\alpha_i(t_0)$ et $\int_0^{t_i}N_i(s^-)d\alpha_i(s)$ étant nuls, on a :

$$N_i(t_i)\alpha_i(t_i)=\int_0^{t_i}\alpha_i(s)dN_i(s)$$

Le processus du score peut alors s'écrire de la manière suivante :

$$U(\beta,t)=\sum_i\left[-\int_0^t Y_i(u)z_i\lambda_c(u,z_i)du+\int_0^t\frac{Y_i(u)z_i\lambda_c(u,z_i)}{\lambda_c(u,z_i)+\lambda_a(a_i+u,x_i)}dN_i(u)\right]$$

$$U(\beta,t)=\sum_i\left[-\int_0^t\frac{Y_i(u)z_i\lambda_c(u,z_i)}{\lambda_c(u,z_i)+\lambda_a(a_i+u,x_i)}\left[\lambda_c(u,z_i)+\lambda_a(a_i+u,x_i)\right]du+\int_0^t\frac{Y_i(u)z_i\lambda_c(u,z_i)}{\lambda_c(u,z_i)+\lambda_a(a_i+u,x_i)}dN_i(u)\right]$$

$$U(\beta,t)=\sum_i\left[\int_0^t\frac{Y_i(u)z_i\lambda_c(u,z_i)}{\lambda_c(u,z_i)+\lambda_a(a_i+u,x_i)}\left[dN_i(u)-\lambda_c(u,z_i)du-\lambda_a(a_i+u,x_i)du\right]\right]$$

$$U(\beta,t)=\sum_i\left[\int_0^t\frac{z_i\lambda_c(u,z_i)}{\lambda_c(u,z_i)+\lambda_a(a_i+u,x_i)}\left[Y_i(u)dN_i(u)-Y_i(u)\lambda_c(u,z_i)du-Y_i(u)\lambda_a(a_i+u,x_i)du\right]\right]$$

$$U(\beta,t)=\sum_i\left[\int_0^t\frac{z_i\lambda_c(u,z_i)}{\lambda_c(u,z_i)+\lambda_a(a_i+u,x_i)}\left[dN_i(u)-Y_i(u)\lambda_c(u,z_i)du-Y_i(u)\lambda_a(a_i+u,x_i)du\right]\right]$$

$$U(\beta,t)=\sum_i\left[\int_0^t\frac{z_i\lambda_c(u,z_i)}{\lambda_c(u,z_i)+\lambda_a(a_i+u,x_i)}dM_i(u)\right]$$

$$U(\beta,t)=\sum_i\left[\int_0^t\frac{\partial\log\lambda_o(u,z_i)}{\partial\beta}dM_i(u)\right]$$

Dans le cadre des taux de mortalité en excès, le processus du score peut donc également s'écrire comme une transformée de martingale avec :

$$h_i(u,x)=\frac{Y_i(u)z_i\lambda_c(u,z_i)}{\lambda_c(u,z_i)+\lambda_a(a_i+u,x_i)}$$

Appendix 6 : Résidus de martingale cumulés pour la vérification de la forme fonctionnelle de la covariable d'intérêt pour un modèle paramétrique du taux en excès

Pour alléger l'écriture, nous supposerons qu'il n'y a qu'une seule covariable incluse dans le modèle.

Les résidus de martingales cumulées pour la covariable z s'expriment de la manière suivante :

$$M_{cum}(x,\beta) = \sum_{i=1}^{n} \int_0^{\tau} I(z_i \le x) dM_i(u)$$

avec $M_i(t) = N_i(t) - \int_0^t Y_i(u)\lambda_c(u)du - \int_0^t Y_i(u)\lambda_a(u)du$

De même que pour les processus du score, les approximations utilisées dans le cadre de la survie globale peuvent être utilisée dans le cadre de la survie nette. Le processus est similaire que l'on se place dans le cadre de la survie globale ou dans le cadre de la survie nette. Le changement réside dans l'expression des processus gaussiens simulés approximant la distribution du processus sous l'hypothèse nulle.

En utilisant le développement de Taylor sur la formule ci-dessus, on obtient le développement suivant :

$$n^{1/2} M_{cum}(x,\hat{\beta}) \approx n^{1/2} M_{cum}(x,\beta_0) + n^{1/2}(\hat{\beta} - \beta_0) \sum_{i=1}^{n} \int_0^{\tau} I(z_i \le x) \frac{\partial d\hat{M}_i(t)}{\partial \beta_0}$$

En utilisant l'approximation faite dans la sous-partie précédente sur le processus du score,

$$n^{1/2}\left(\hat{\beta}-\beta_0\right)\approx n^{-1/2}U(\beta_0;\tau)\times nI\left(\hat{\beta};\tau\right)^{-1}$$

et en développant le terme possédant une intégrale,

$$\frac{\partial d\hat{M}_i(t)}{\partial\beta_0}=\frac{\partial dN_i(t)}{\partial\beta_0}-\frac{\partial d\Lambda_{ci}(t,z_i,\beta_0)}{\partial\beta_0}-\frac{\partial d\Lambda_{ai}(t,x_i,\beta_0)}{\partial\beta_0}$$
$$\frac{\partial d\hat{M}_i(t)}{\partial\beta_0}=-\frac{\partial d\Lambda_{ci}(t,z_i,\beta_0)}{\partial\beta_0}$$

avec

$$-\frac{\partial d\Lambda_{ci}(t,z_i,\beta_0)}{\partial\beta_0}=-\frac{\partial\exp(\beta_0 z_i)d\Lambda_0(t)}{\partial\beta_0}$$
$$-\frac{\partial d\Lambda_{ci}(t,z_i,\beta_0)}{\partial\beta_0}=-z_i\exp(\beta_0 z_i)d\Lambda_0(t)$$

Nous obtenons,

$$n^{1/2}M_{cum}\left(x,\hat{\beta}\right)\approx n^{1/2}M_{cum}(x,\beta_0)-n^{-1/2}U(\beta_0;\tau)\times nI\left(\hat{\beta};\tau\right)^{-1}\sum_{i=1}^{n}\int_0^\tau I(z_i\le x)z_i\exp(\beta_0 z_i)d\Lambda_0(t)$$
$$n^{1/2}M_{cum}\left(x,\hat{\beta}\right)\approx n^{1/2}M_{cum}(x,\beta_0)-n^{-1/2}U(\beta_0;\tau)\times nI\left(\hat{\beta};\tau\right)^{-1}\times J(\tau,z,\beta_0)$$

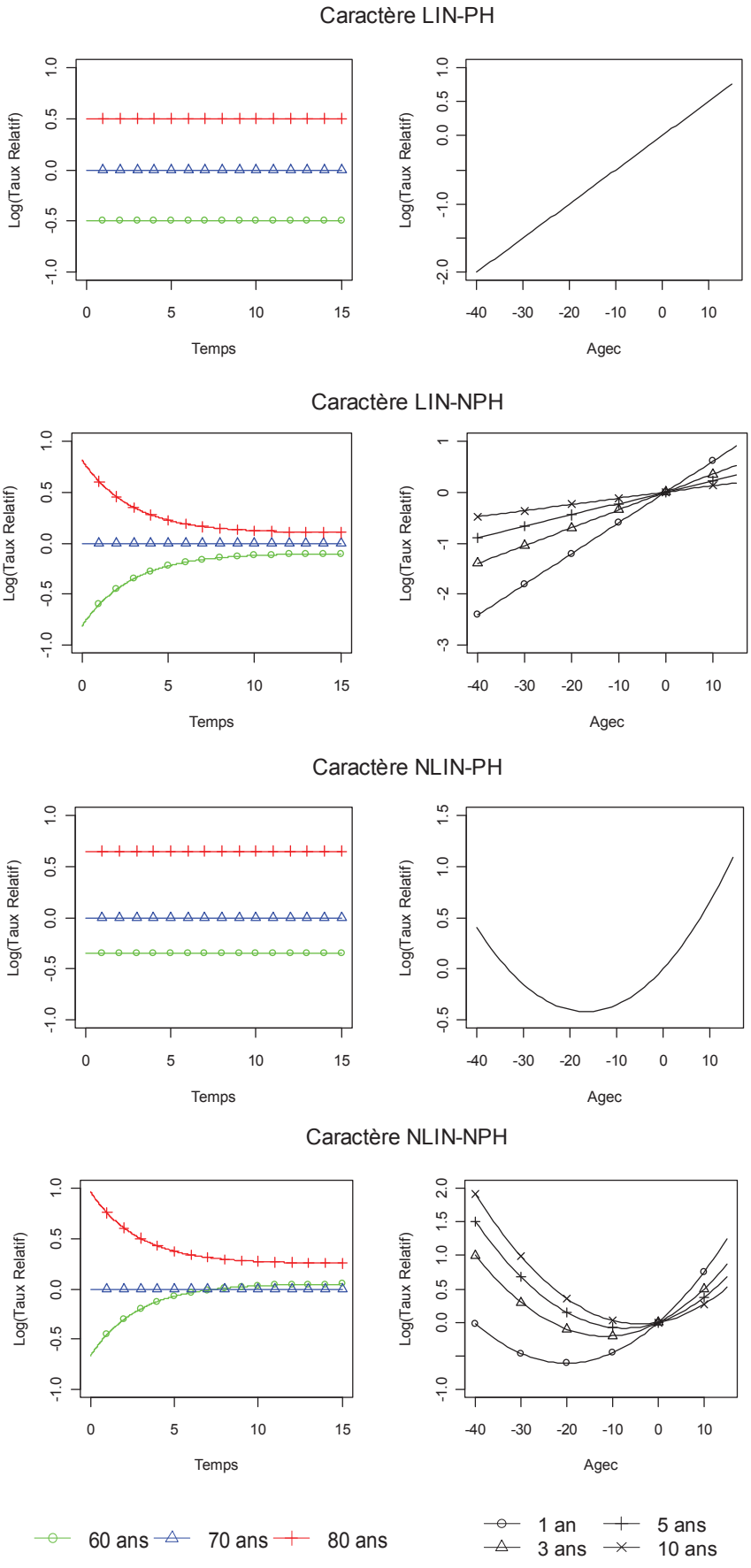
La distribution de $n^{1/2} M_{cum}\left(x, \hat{\beta}\right)$ est donc asymptotiquement équivalente à la distribution du processus W_z, tel que:

$$W_z(x) \approx^{1/2} M_{cum}(x, \beta_0) - n^{-1/2} U(\beta_0; \tau) \times nI\left(\hat{\beta}; \tau\right)^{-1} \times J(\tau, z, \beta_0)$$

Les différences existantes entre le cadre de la survie globale et le cadre de la survie nette pour le test de la fonction de lien sont identiques aux différences existantes pour le test de la forme fonctionnelle.

Appendix 7 : Illustration des données simulées

Figure III.17. Illustration du caractère (non)-linéaire-(non)-proportionnel des données simulées



Chapitre IV

Survie nette conditionnelle - Application épidémiologique au cancer du côlon

IV.1. Introduction

Dans les études de survie, la probabilité de survie nette est traditionnellement reportée depuis le diagnostic. Hors on observe quasiment toujours un taux de mortalité instantané en excès variable au cours du temps : ce taux n'étant pas constant, il s'ensuit que la probabilité de survivre à 5 ans (par exemple) n'est pas la même selon que l'on se situe tout de suite après le diagnostic ou au-delà ; les estimations de survie classique ne sont donc plus appropriée.

C'est pourquoi la survie nette conditionnelle a été récemment introduite dans la littérature. Elle consiste à estimer la survie d'un groupe de patients après le temps *t*, conditionnellement au fait d'avoir survécu jusqu'au temps *t*. L'analyse de la survie nette conditionnelle rejoint complétement celle de la dynamique du taux de mortalité instantané en excès. Après avoir survécu 1, 3 ou 5 ans, quel est le taux de mortalité en excès lié au cancer encore subi par les patients survivants ? Elle permet également de restituer l'effet de certains facteurs sur le taux de mortalité en excès et d'évaluer par exemple si l'un de ces facteurs, connu au moment du diagnostic, a toujours un effet *x* années au-delà, c'est-à-dire d'évaluer si leur effet est constant ou non au cours du temps.

Cet indicateur n'est donc finalement qu'une autre façon de restituer la dynamique du taux de mortalité en excès. Toutefois, son interprétation semble plus aisée pour les épidémiologistes et les cliniciens, comme en témoigne l'augmentation récente des publications le rapportant. Disposer de probabilités de survie mises à jour au fur et à mesure que le temps s'écoule est en effet une information appréciée par les cliniciens : cela permet d'adapter leurs stratégies de suivi au fur et à mesure que le temps s'écoule. De même, cet indicateur (tout comme la dynamique du taux de mortalité en excès) pourrait être particulièrement utilisé dans le contexte de l'assurabilité des patients. En effet, dès lors qu'ils souhaitent faire un emprunt, les patients atteints de cancer sont le plus souvent considérés comme ayant un risque de santé aggravé et à ce titre, soit l'emprunt leur est refusé, soit il est accepté avec des surprimes conséquentes. Fournir des probabilités de survie conditionnelle et

montrer que dans un grand nombre de cas celle-ci augmente avec le temps devraient permettre aux patients d'emprunter sans surprime dès lors qu'un certain temps s'est écoulé depuis le diagnostic.

D'une manière générale, le nombre de patients survivant à un cancer augmentant, il serait judicieux de fournir d'avantage de résultats en termes de survie nette conditionnelle. On note tout de même une forte augmentation de son utilisation dans les analyses de survie de ces dernières années (pendant la période 2010-2013, quarante-et-un articles avaient les termes « conditional survival » dans leur titre dans la base PubMed alors qu'ils n'étaient que trois dans la période 1998-2001).

Le stade au diagnostic étant un facteur pronostique indiscutable pour l'estimation de la survie, l'objectif sera d'étudier si l'impact du stade au diagnostic sur la dynamique du taux de mortalité en excès associé au cancer du côlon entre 0 et 15 ans de suivi est le même tout au long du suivi. Cette étude porte sur le cancer du côlon qui, avec 42152 nouveaux cas observés en 2012 en France (dont 55% d'hommes), représente le troisième cancer le plus fréquent chez les hommes et le deuxième chez les femmes. Malgré le fait que l'incidence et la mortalité décroissent depuis quelques années, la mortalité liée à ce cancer reste encore l'une des plus élevées parmi les différentes localisations de cancer.

IV.2. Matériels et Méthodes

IV.2.1. Matériels

Cette étude repose sur des données issues d'études Haute-Résolution menées par le réseau Francim. Les données Hautes-Résolutions (HR) contiennent des informations collectées en phase initiale telles que des données concernant le patient (sexe, date de naissance, lieu de résidence), et des données concernant le cancer (localisation, stade au diagnostic, traitement initial). Ces données HR contiennent également des informations sur l'évolution clinique des cas pendant cinq années après le diagnostic (type de la (les) reprises évolutives, traitements secondaires, lieu des traitements secondaires…). Au-delà de cette période, seul le statut vital est enregistré.

Table VI.1. Effectifs par âge et par stade des patients atteints du cancer du Colon diagnostiqués en 1990 dans 7 registres français

	[15 ; 45]	**[45 ; 55]**	**[55 ; 65]**	**[65 ; 75]**	**[75 ; ++]**	**Total**
Stade 1	18 (6.0 %)	25 (8.3 %)	69 (22.9%)	84 (27.9 %)	105 (34.9 %)	301 (17.8 %)
Stade 2	16 (3.2 %)	24 (4.8 %)	94 (18.6 %)	169 (33.5 %)	202 (40.0 %)	505 (29.9 %)
***Stade 2a**[1]*	*8 (4.8 %)*	*10 (6.0 %)*	*42 (25.1 %)*	*56 (33.5 %)*	*51 (30.5 %)*	*167 (9.9 %)*
***Stade 2b**[2]*	*4 (1.8 %)*	*9 (4.1 %)*	*31 (14.3 %)*	*61 (28.1 %)*	*112 (51.6 %)*	*217 (12.8 %)*
***Stade 2c**[3]*	*4 (3.3%)*	*5 (4.1%)*	*21 (17.4%)*	*52 (43.0%)*	*39 (32.2%)*	*121 (7.2 %)*
Stade 3	9 (2.5 %)	34 (9.6 %)	72 (20.3 %)	95 (26.8 %)	145 (40.8 %)	355 (21.0 %)
Stade 4	8 (1.9 %)	37 (8.6 %)	77 (17.9 %)	125 (29.1 %)	183 (42.6 %)	430 (25.4 %)
Non stadés	1 (1.0 %)	2 (2.0 %)	16 (16.2 %)	19 (19.2 %)	61 (61.6 %)	99 (5.9 %)
Tous stades confondus	52 (3.1 %)	122 (7.2 %)	328 (19.4 %)	492 (29.1 %)	696 (41.2 %)	1690 (100 %)

*[1] Stade 2a : Stade 2 avec plus de 8 ganglions examinés. *[2] Stade 2b : Stade 2 avec moins de 8 ganglions examinés. *[3] Stade 2c : Stade 2 avec un nombre de ganglions examinés manquants.

Les données utilisées sont les données HR du cancer du côlon recueillies auprès de sept registres du réseau Francim (Calvados, Côte-d'Or, Doubs, Hérault, Saône-et-Loire, Somme et Tarn) dont les patients ont été diagnostiqués en 1990 et suivis jusqu'au 31/12/2007 (il s'agit d'un échantillon des cas diagnostiqués en 1990) : cela représente 1690 cas dont

55.33% d'hommes et 44.67% de femmes. Les effectifs par âge et par stade sont présentés dans le tableau IV.1. La majorité des patients atteints de cancer du côlon ont plus de 65 ans (plus de 70%). Les patients présentant un cancer de stade 2 sont les plus nombreux et représentent presque 30% de la population, puis viennent les stades 4 (25.4%), les stades 3 (21%) et les stades 1 (18%). Le stade n'est pas connu pour 6% des patients. Les stades 2 peuvent être divisés en trois sous-groupes selon leur nombre de ganglions examinés, que l'on appellera stade 2a (plus de 8 ganglions examinés), stade 2b (mois de 8 ganglions examinés) et stade 2c (nombre de ganglions manquants). Lorsque le nombre de ganglions examinés est inférieur à 8, la stadification n'est pas supposée très fiable, il est donc attendu que les stades 2b soit un mélange de patients diagnostiqués en stade 2 et en stade 3 avec un moins bon pronostic que les patients diagnostiqués en stade 2a qui sont de « vrais » stade 2.

Dans cette étude, nous nous focaliserons particulièrement sur les patients diagnostiqués en stade 2, en stade 2a et en stade 3. Les variables pronostiques que nous étudierons sont l'*âge* du patient au diagnostic ainsi que le *stade au diagnostic*. La covariable *sexe* ne sera pas prise en compte du fait que les résultats obtenus pour les hommes et les femmes sont très proches pour ce site. L'analyse sera donc faite tous sexes confondus.

Table IV.2. Description du nombre de décès et de perdus de vue par stade et par classe d'âge

		Nombre de décès/Perdus de vue à différents temps après le diagnostic				
Stade	Classe d'âge	Temps depuis le diagnostic (années)				
		1	3	5	10	15
Tous stades (1690)		450/9	793/13	965/18	1167/23	1325/31
	[26 ;44] (52)	5/2	10/2	13/2	16/5	18/5
	[45 ;54] (122)	21/0	47/1	55/1	69/1	74/5
	[55 ;64] (328)	65/1	130/1	157/2	186/2	208/3
	[64 ;75] (492)	114/0	207/0	259/2	312/3	371/5
	[75 ;++] (696)	245/6	402/9	481/11	584/12	654/13
Stade 1 (301)		17/2	47/3	68/6	117/9	168/12
	[26 ;44] (18)	0/1	0/1	0/1	1/4	3/4
	[45 ;54] (25)	0/0	0/0	0/0	2/0	4/1
	[55 ;64] (69)	1/0	7/0	10/1	16/1	23/2
	[64 ;75] (84)	2/0	12/0	15/1	26/1	43/2
	[75 ;++] (105)	14/1	28/2	43/3	72/3	95/3
Stade 2 (505)		56/2	123/3	204/4	295/5	364/7
	[26 ;44] (16)	0/0	2/0	3/0	5/0	5/0
	[45 ;54] (24)	2/0	5/1	7/1	10/1	12/2
	[55 ;64] (94)	2/0	10/0	25/0	36/0	47/0
	[64 ;75] (169)	17/0	34/0	59/1	84/2	114/3
	[75 ;++] (202)	35/2	72/2	110/2	160/2	186/2
Stade 2a (167)		15/0	30/1	49/1	75/1	102/1
	[26 ;44] (8)	0/0	0/0	1/0	1/0	1/0
	[45 ;54] (10)	2/0	3/1	3/1	3/1	3/1
	[55 ;64] (42)	0/0	2/0	7/0	14/0	20/0
	[64 ;75] (56)	4/0	6/0	15/0	21/0	33/0
	[75 ;++] (51)	9/0	16/0	23/0	36/0	45/0
Stade 3 (355)		67/2	179/4	219/5	264/5	290/7
	[26 ;44] (9)	0/1	0/1	2/1	2/1	2/1
	[45 ;54] (34)	3/0	9/0	13/0	21/0	21/2
	[55 ;64] (72)	10/1	35/1	40/1	49/1	51/1
	[64 ;75] (95)	13/0	43/0	59/0	70/0	79/0
	[75 ;++] (145)	41/0	92/2	105/3	122/3	137/3
Stade 4 (430)		280/1	395/1	414/1	419/1	422/1
	[26 ;44] (8)	5/0	8/0	8/0	8/0	8/0
	[45 ;54] (37)	16/0	33/0	35/0	36/0	36/0
	[55 ;64] (77)	49/0	72/0	75/0	75/0	76/0
	[64 ;75] (125)	78/0	111/0	119/0	121/0	122/0
	[75 ;++] (183)	132/1	171/1	177/1	179/1	180/1

La table IV.2 décrit le nombre de décès et de perdus de vue à 1, 3, 5, 10 et 15 ans de suivi après le diagnostic. Après 15 ans de suivi, 78.4% des patients sont décédés et 1.8% ont été perdus de vue. Le pourcentage de décès augmente avec l'âge ; en effet, 34.6% des patients diagnostiqués dans la classe la plus jeune décèdent contre 94.0% dans la classe d'âge la plus âgée. Nous pouvons remarquer que cette tendance s'observe également au sein de chaque stade excepté pour le stade 4 pour lequel le taux de mortalité est tellement fort que le pourcentage de décès est équivalent pour toutes les classes d'âge et proche de 100%. Les quantiles des temps de décès diminuent avec la gravité du stade. En ce qui concerne le 50ème et le 95ème percentile des temps de décès, ils sont respectivement de 8.03 ans et 16.88 ans pour

les stades 1, 4.66 ans et 15.25 ans pour les stades 2, 2.32 ans et 12.98 ans pour les stades 3 et 0.61 an et 3.25 ans pour les stades 4. Le fait que la plupart des décès se situent en début de suivi, particulièrement pour les stades 4, induit un manque d'information à long terme qui va probablement rendre difficile certaines estimations.

IV.2.2. Méthodes

IV.2.2.a. Estimation de la survie nette

La survie nette sera estimée à l'aide de modèles paramétriques de régression du taux de mortalité en excès. Pour chaque stade analysé séparément, le modèle suivant [Remontet, 2007] a été utilisé :

$$\log[\lambda_c(t, agec)] = f(t) + h(t) \times agec + g(agec)$$

où la fonction *f*, représentant le taux de base, est un spline cubique à 2 nœuds fixés à 1 et 5 ans, la fonction *h*, représentant la partie dépendante du temps de l'effet de l'âge, est choisie parmi quatre fonctions candidates (un spline cubique à un nœud fixé à 1 an, une fonction cubique, une fonction quadratique et une fonction linéaire) et la fonction *g*, représentant la partie linéaire ou non-linéaire de l'effet de l'âge, est choisie parmi quatre fonctions candidates (un spline cubique avec un nœud à l'âge centré (âge - 70) égal 0, une fonction cubique, une fonction quadratique et une fonction linéaire). La meilleure combinaison de ces trois fonctions est choisie à l'aide du critère d'Akaike [Akaike, 1973]. Cette stratégie a été établie pour l'analyse des données des stades 2, des stades 2a et des stades 3 (table IV.3). Les patients diagnostiqués en stade 1 n'apparaissent pas dans l'analyse pour cause de problème de convergence liée au fait qu'il y a peu de décès dû au cancer, surtout après 7 ans de suivi. De même, les patients diagnostiqués en stades 4 n'apparaissent pas dans l'analyse du fait de leur mortalité précoce ; 95% des décès ont lieu pendant les quatre premières années. Les patients diagnostiqués en stade 2a apparaissent dans l'analyse car du fait qu'ils aient plus de 8 ganglions examinés, le niveau du stade au diagnostic qui leur a été affecté est supposé fiable ;

ils représentent donc potentiellement les « vrais » stades 2. Le modèle retenu pour chacun des stades sont décrits dans la table IV.4.

Table IV.3. Fonctions candidates pour modéliser le taux en excès des patients en stade 2, stade 2a et stade 3

Stade	Fonctions candidates		
	f	*h*	*g*
Stade 2, 2a, 3	Spline cubique à 2 nœuds (fixés à 1 et 5 ans)	Spline cubique à 1 nœud (fixé à 1 an) Une fonction cubique Une fonction quadratique Une fonction linéaire	Spline cubique à 1 nœud (à l'âge centré égal 0) Une fonction cubique Une fonction quadratique Une fonction linéaire

Table IV.4. Fonctions candidates retenues pour modéliser le taux en excès les stades 2, 2a et 3

Stade	Fonctions candidates		
	f	*h*	*g*
Stade 2	Spline cubique à 2 nœuds (fixés à 1 et 5 ans)	Spline cubique à 1 nœud (fixé à 1 an)	Une fonction linéaire
Stade 2a	Spline cubique à 2 nœuds (fixés à 1 et 5 ans)	Une fonction quadratique	Une fonction linéaire
Stade 3	Spline cubique à 2 nœuds (fixés à 1 et 5 ans)	Spline cubique à 1 nœud (fixé à 1 an)	Spline cubique à 1 nœud (à l'âge centré égal 0)

La survie nette, S_c, a ensuite été estimée à l'aide de la relation usuelle qui lie la survie et taux de mortalité :

$$S_c(t, agec) = \exp\left(-\int_0^t \lambda_c(t, agec)\right)$$

Son intervalle de confiance à 95% a été calculé à l'aide de la méthode de Monte-Carlo. Soit β^* le vecteur des paramètres du modèle estimés à l'aide du maximum de vraisemblance. A l'aide de la loi normale multivariée, 1000 vecteurs de moyenne β^* et de variance la matrice de variance-covariance de β^* ont été générés. A partir de ces 1000 générations, l'intervalle de

confiance a été calculé en prenant le 2.5ème et 97.5ème percentile des 1000 survies nette estimées.

IV.2.2.b. Estimation de la survie nette conditionnelle

La survie nette conditionnelle, notée $SC_c(y|x)$, est la probabilité qu'un patient soit vivant y années après le diagnostic sachant avoir déjà survécu x années ($x < y$). Cette quantité se mesure de la manière suivante :

$$SC_c(y|x) = \frac{S_c(y)}{S_c(x)}$$

Ce qui est équivalent à :

$$SC_c(y|x) = \exp\left(-\int_x^y \lambda_c(u)du\right)$$

Dans notre étude, nous nous intéresserons à la survie nette conditionnelle à ($y = x+5$) ans après avoir survécu x années après le diagnostic, $x \in \{1,\ldots,10\}$.

L'intervalle de confiance de la survie nette conditionnelle est calculé à l'aide de la méthode de Monte-Carlo de la même manière que pour la survie nette. A partir de ces 1000 générations, l'intervalle de confiance a été calculé en prenant le 2.5ème et 97.5ème percentile des 1000 survies nettes conditionnelles estimées.

IV.2.2.c. Estimation des taux relatifs

Les taux relatifs, notes *TR*, ont été estimés en faisant le rapport du taux estimé pour le stade 3 et du taux estimé pour le stade 2 ou stade 2a pris comme référence :

$$TR(t) = \frac{\lambda_{c,Stade3}(t)}{\lambda_{c,Stade2}(t)}$$

Les intervalles de confiance de ces taux relatifs ont été estimés à l'aide de la méthode de Monte-Carlo comme la survie nette et la survie nette conditionnelle. A partir des 1000 vecteurs de moyenne β^* et de variance la matrice de variance-covariance de β^* générés pour l'analyse des stades 2/2a et 3, 1000 taux relatifs ont été générés comparant les stades 3 aux stades 2/2a. L'intervalle de confiance des taux relatifs a été calculé en prenant le 2.5^{ème} et 97.5^{ème} percentile des 1000 taux relatifs estimés.

IV.3. Résultats

Les patients de cette étude étant tous diagnostiqués en 1990 et suivis jusqu'au 31 Décembre 2007, la survie nette a pu être estimée jusqu'à 15 ans de suivi. La table IV.5 présente les résultats de la survie nette jusqu'à 15 ans de suivi pour chaque stade. La figure IV.1 présente les courbes de survie obtenues pour chaque stade ainsi que l'évolution des taux de mortalité instantanés obtenus également pour chaque stade avec la méthode de Remontet, et le modèle d'Estève [Estève, 1990]. La figure IV.2 fournit le graphe de l'évolution des taux de mortalité au cours du temps par classe d'âge en fonction du stade. Puis sont ensuite présentés les résultats en terme de survie nette conditionnelle avec la table IV.5 et la figure IV.3 qui présentent les résultats de la survie nette conditionnelle pour chaque stade, à (x+5) ans sachant que le patient a déjà survécu x années depuis le diagnostic. La figure IV.4 et la figure IV.5 représentent les taux relatifs, rapport du taux de mortalité en excès des patients diagnostiqués en stade 2 ou stade 2a avec le taux de mortalité en excès des patients diagnostiqués en stade 3, pour tous âges confondus et par classe d'âge.

IV.3.1. Survie nette

Pour les patients diagnostiqués en stade 2, la survie nette tous âges confondus diminue de 93.5% (à un an de suivi) à 56.2% (à 15 ans de suivi) (table IV.5). Nous pouvons voir à l'aide de la figure IV.1 que la survie nette de ces patients ne semble pas atteindre de palier. Ceci est en cohérence avec les taux de mortalité instantanés qui ne s'annulent pas, même à long terme (figure IV.1). Cependant, lorsque nous regardons les résultats par classe d'âge, nous pouvons voir que les taux de mortalité s'annule pour les âges inférieur à 55 ans. Cette tranche d'âge de la population représentant un faible pourcentage, elle a peu d'impact sur le taux de mortalité marginal. La survie nette décroit avec l'âge. Elle est comprise entre 62.7% (classe d'âge [75 ;++]) et 69.5% (classe d'âge [15 ;45]) à 10 ans de suivi et entre 47.5% (classe d'âge [75 ;++]) et 67.2% (classe d'âge [15 ;45]) à 15 ans de suivi. En observant l'allure des taux de mortalité au cours du temps par classe d'âge (figure IV.2), nous pouvons voir que la différence de mortalité entre les classes d'âges se joue principalement au cours de la première année de suivi après le diagnostic.

Pour les patients diagnostiqués en stade 2a, la survie nette tous âges confondus diminue de 95.1% (à un an de suivi) à 68.9% (à 15 ans de suivi) (table IV.5). Tout comme pour les stades 2, nous pouvons voir à l'aide de la figure IV.1 que la survie nette des patients diagnostiqués en stade 2a ne semble pas atteindre de palier. Ceci est en cohérence avec les taux de mortalité instantanés qui ne s'annulent pas, même à long terme (figure IV.1). Lorsque nous regardons les résultats par classe d'âge, nous pouvons voir que les taux de mortalité ne s'annule pas pour les âges compris dans l'intervalle suivant [55; ++], qui représentent plus de 80% de la population. Jusqu'à environ 10 ans de suivi, la survie croît avec l'âge. A 10 ans de suivi, la survie nette est comprise entre 79.4% (classe d'âge [15 ;45]) et 80.6% (classe d'âge [75 ;++]) et à 15 ans de suivi, elle est comprise entre 58.4% (classe d'âge [75 ;++]) et 78.0% (classe d'âge [15 ;45]).

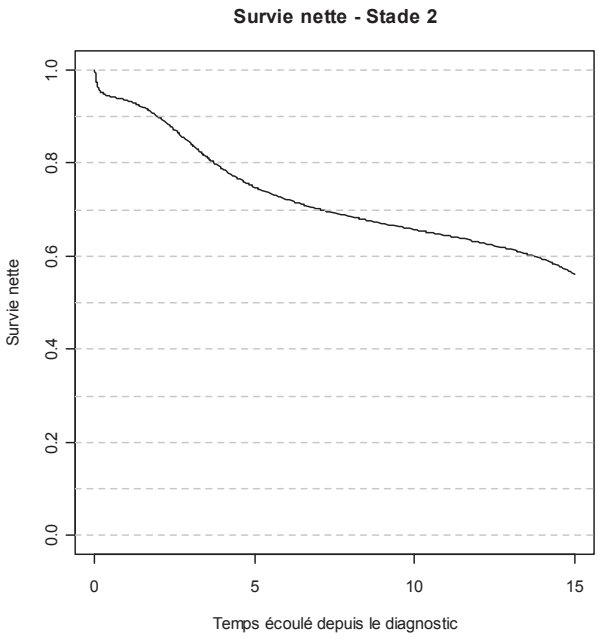
En ce qui concerne les patients diagnostiqués en stade 3, nous pouvons observer que la survie nette diminue de 82.5 % (à un an de suivi) à 30.8% (à 15 ans de suivi) (table IV.5). De même que pour l'analyse des stades 2 et les stades 2a, la survie nette ne semble pas atteindre de palier. Ceci est en cohérence avec le graphe des taux de mortalité instantanés qui ne s'annulent pas, même à long terme (figure IV.1). Cependant, le manque d'information en fin de suivi rend difficile l'interprétation. En ce qui concerne les résultats par classe d'âge, la

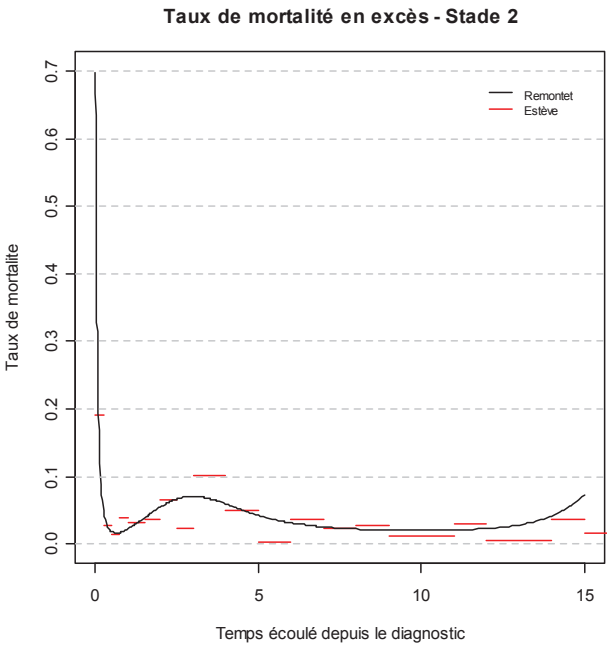
survie nette est comprise entre 33.0% (classe d'âge [75 ;++]) et 78.0% (classe d'âge [15 ;45]) à 10 ans de suivi et entre 19.5% (classe d'âge [75 ;++]) et 78.0% (classe d'âge [15 ;45]) à 15 ans de suivi. En observant l'évolution des taux de mortalité au cours du temps par classe d'âge (figure IV.2), nous pouvons voir que le taux de mortalité atteint une valeur maximale entre 0 et 5 ans. Globalement, plus la classe d'âge est élevée, plus la taux de mortalité est élevé, excepté pour les classes d'âge [55 ;65] et [65 ;75] qui ont un profil assez proches.

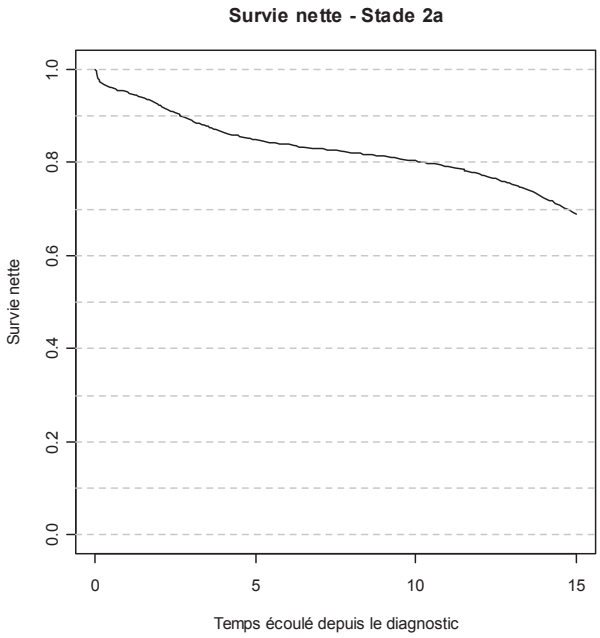
Table IV.5. Survie nette à 1, 3, 5, 10 et 15 ans

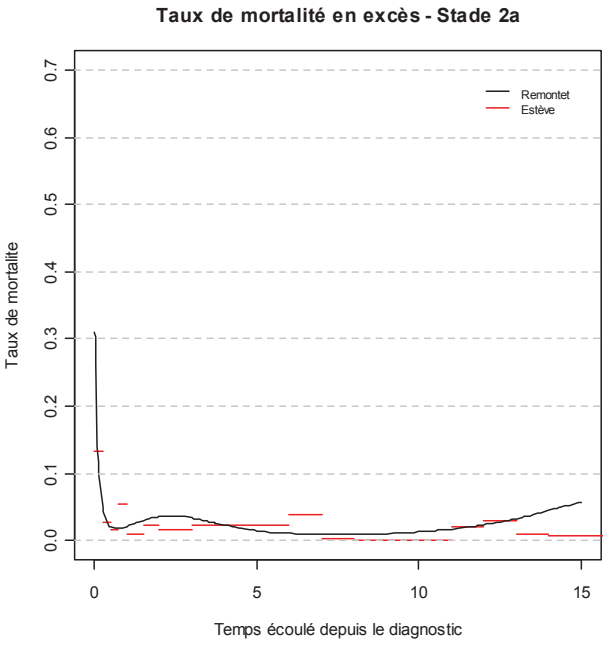
	Survie nette (années après le diagnostic)				
	1	**3**	**5**	**10**	**15**
Stade 2	**93.5 [89.8 ;95.4]**	**84.2 [78.3 ; 87.4]**	**74.9 [67.7 ; 79.4]**	**65.7 [55.4 ; 71.5]**	**56.2 [41.0 ; 64.0]**
[15 ;45]	97.3 [89.3 ; 99.2]	86.4 [69.6 ; 93.4]	78.6 [59.9 ; 87.5]	69.5 [46.1 ; 80.7]	67.2 [41.4 ; 79.8]
[45 ;55]	97.2 [92.8 ; 98.8]	86.9 [77.8 ; 91.6]	78.5 [67.3 ; 84.8]	69.2 [55.1 ; 77.4]	66.0 [48.5 ; 74.7]
[55 ;65]	96.6 [93.3 ; 98.1]	86.7 [81.4 ; 90.2]	77.7 [71.2 ; 82.0]	68.4 [59.3 ; 74.2]	63.8 [53.2 ; 70.3]
[65 ;75]	95.2 [91.7 ; 97.0]	85.7 [80.4 ; 89.1]	76.3 [69.7 ; 81.1]	67.0 [57.9 ; 73.2]	59.8 [47.4 ; 67.4]
[75 ;++]	89.9 [83.7 ; 93.3]	81.2 [72.3 ; 86.7]	71.6 [60.6 ; 79.2]	62.7 [47.1 ; 71.7]	47.5 [21.3 ; 62.1]
Stade 2a	**95.1 [87.2 ; 97.5]**	**89.1 [76.6 ; 93.5]**	**84.9 [69.8 ; 90.3]**	**80.4 [55.2 ; 87.6]**	**68.9 [41.4 ; 79.4]**
[15 ;45]	94.9 [86.1 ; 97.5]	87.8 [70.1 ; 93.9]	82.9 [54.8 ; 91.9]	79.4 [40.0 ; 90.5]	78.0 [30.3 ; 89.4]
[45 ;55]	95.0 [86.4 ; 97.5]	88.2 [72.9 ; 93.6]	83.6 [65.8 ; 90.7]	79.9 [53.7 ; 88.2]	77.4 [42.0 ; 86.3]
[55 ;65]	95.1 [87.1 ; 97.6]	88.7 [76.1 ; 93.3]	84.4 [70.7 ; 89.9]	80.3 [63.1 ; 87.1]	75.1 [52.5 ; 82.6]
[65 ;75]	95.1 [87.2 ; 97.5]	89.1 [76.8 ; 93.7]	85.0 [70.9 ; 90.9]	80.5 [56.3 ; 88.7]	70.9 [42.1 ; 82.3]
[75 ;++]	95.2 [87.0 ; 97.5]	89.6 [75.2 ; 94.5]	85.8 [62.1 ; 92.7]	80.6 [35.8 ; 91.2]	58.4 [13.0 ; 82.5]
Stade 3	**82.5 [77.4 ; 85.8]**	**55.1 [48.7 ; 60.1]**	**44.0 [37.1 ; 49.1]**	**36.8 [25.7 ; 42.8]**	**30.8 [05.2 ; 37.0]**
[15 ;45]	98.9 [92.6 ; 99.7]	93.0 [67.1 ; 97.4]	86.0 [45.1 ; 94.6]	78.0 [24.8 ; 91.2]	78.0 [0.00 ; 91.2]
[45 ;55]	94.1 [86.7 ; 97.0]	71.7 [57.6 ; 79.9]	55.2 [40.0 ; 65.4]	41.5 [23.1 ; 52.8]	41.3 [0.00 ; 52.3]
[55 ;65]	89.1 [82.8 ; 92.4]	60.4 [50.5 ; 67.7]	45.6 [35.9 ; 53.6]	35.3 [24.6 ; 43.9]	34.8 [0.00 ; 42.9]
[65 ;75]	85.0 [79.6 ; 89.1]	56.6 [47.1 ; 63.5]	45.1 [36.0 ; 53.5]	38.0 [26.6 ; 47.3]	36.7 [0.00 ; 45.7]
[75 ;++]	73.8 [65.4 ; 80.0]	45.1 [35.5 ; 54.2]	37.3 [26.0 ; 47.2]	33.0 [16.4 ; 44.0]	19.5 [03.4 ; 32.8]

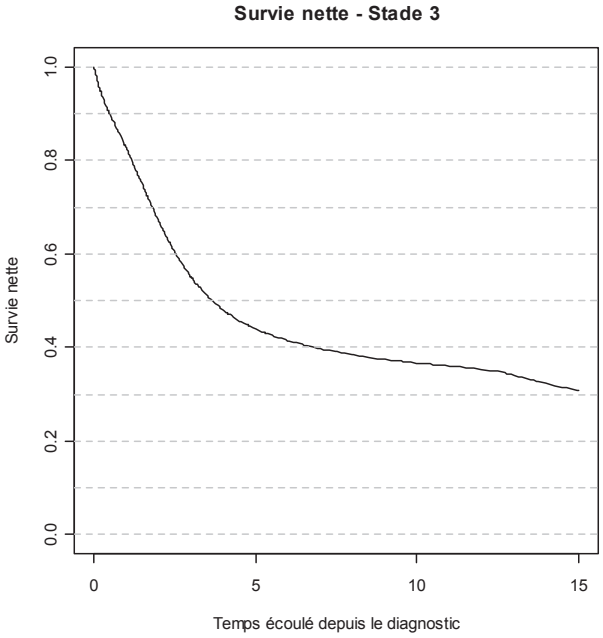
Figure IV.1. Survie et Taux de mortalité au cours du temps pour les stades 2, 2a et 3











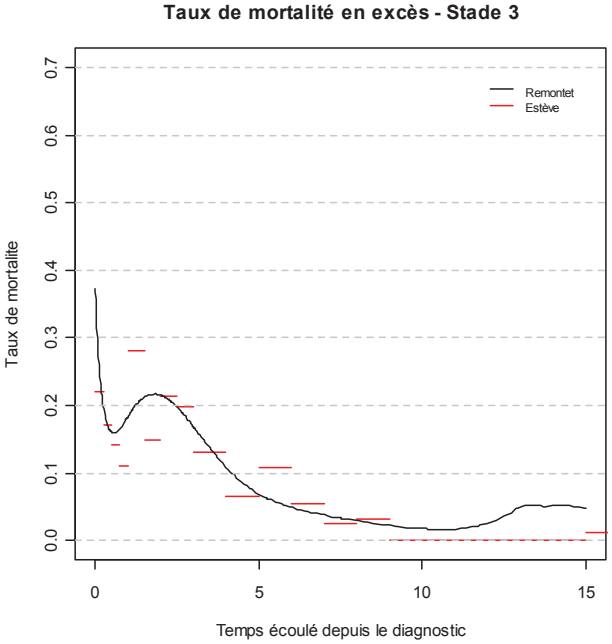
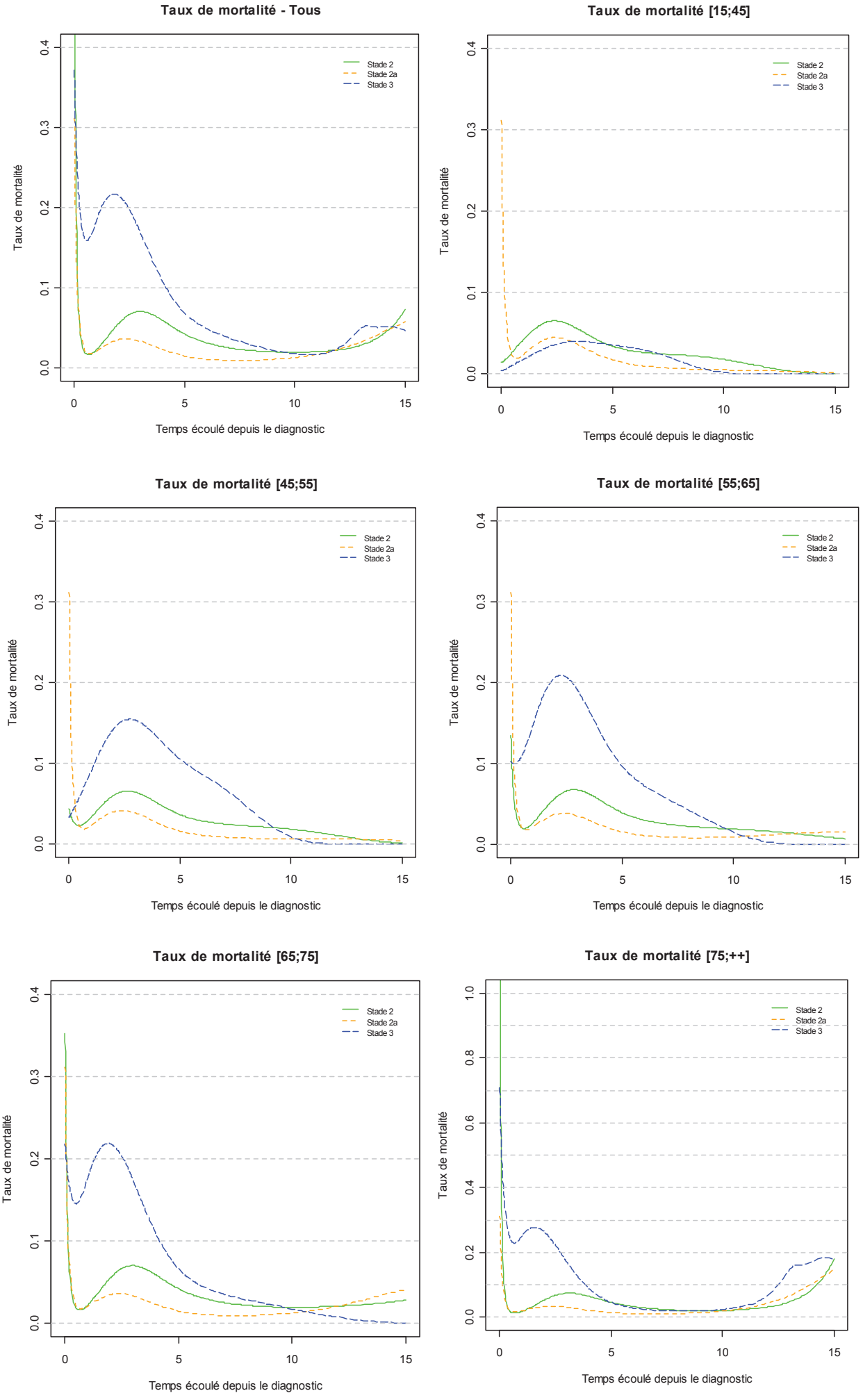


Figure IV.2. Taux de mortalité au cours du temps par classe d'âge en fonction du stade



IV.3.2. Survie nette conditionnelle

Les courbes de survie nette conditionnelle pour chaque stade (figure IV.3) se lisent ainsi : chaque point représente la probabilité de survie nette à (x+5) ans conditionnellement au fait d'avoir survécu x années depuis le diagnostic. Quand x = 0, cela équivaut à la survie à 5 ans habituelle, c'est-à-dire la survie mesurée depuis le diagnostic. Quand x = 6 ans par exemple, l'ordonnée correspond à la probabilité d'être en vie à 11 ans après le diagnostic sachant qu'on est vivant à 6 ans. Nous pouvons observer que cette survie nette conditionnelle augmente avec le temps et ceci jusqu'à 8 ans pour les stades 2, jusqu'à 5 ans pour les stades 2a et jusqu'à 7 ans pour les stades 3. Cette observation se fait l'écho de celle faite à propos de l'évolution du taux en excès : celui-ci diminuant avec le temps (figures IV.1 et IV.2 - tous âges), la survie nette conditionnelle augmente. Il existe cependant une légère décroissance de cette survie conditionnelle à partir de 5, 7 ou 8 ans après le diagnostic selon le stade (là encore cohérente avec les graphes d'évolution du taux en excès des figures IV.1 et IV.2 - tous âges). Cette décroissance n'était pas attendue et reste d'interprétation délicate comme cela sera discuté plus loin. Nous pouvons observer que plus le temps s'écoule, plus l'écart entre les courbes associées aux différents stades diminue jusqu'à devenir faible au-delà de 6 ans.

Quand cette évolution est examinée par classe d'âge (table IV.6), on ne voit plus de décroissance de la survie nette conditionnelle sauf chez les personnes âgées de plus de 75 ans pour les stades 2 et 3 et sauf pour les patients âgées de 55 ans ou plus pour les stades 2a. Cela est tout à fait cohérent avec les graphes illustrant les taux en excès par classe d'âge (figure IV.2). Les classes d'âge les plus âgées, et particulièrement la classe d'âge [75 ; ++], ont donc un impact fort sur l'estimation tous âges mais on note la très faible précision des estimations dans cette classe d'âge, surtout à long terme. Pour les stades 2 et 2a, la survie nette conditionnelle à (x+5) ans atteint le seuil de 95%, seuil au-dessus duquel il est suggéré que l'excès de mortalité devient négligeable [VanSteenBergen, 2013], entre 5 et 10 ans suivant le diagnostic, chez les patients de moins 55 ans. Dans cette tranche d'âge, l'excès de mortalité est alors considéré comme négligeable.

On note également en ce qui concerne les stades 2, que la survie nette conditionnelle est relativement proche entre les différentes classes d'âge après avoir survécu 1, 3 ou 5 ans après le diagnostic. Au-delà, la survie nette conditionnelle diminue avec l'âge.

En ce qui concerne les stades 3, on observe globalement les mêmes phénomènes que précédemment mais les estimations de la survie à (x+5) ans après avoir survécu x = 10 ans

sont inexploitables tant l'imprécision est grande. Il faut réaliser que ces estimations utilisent des données de suivi jusqu'à 15 ans qui deviennent dans ce cas rares. On note tout de même que la survie conditionnelle semble diminuer ici avec l'âge dès le début du suivi.

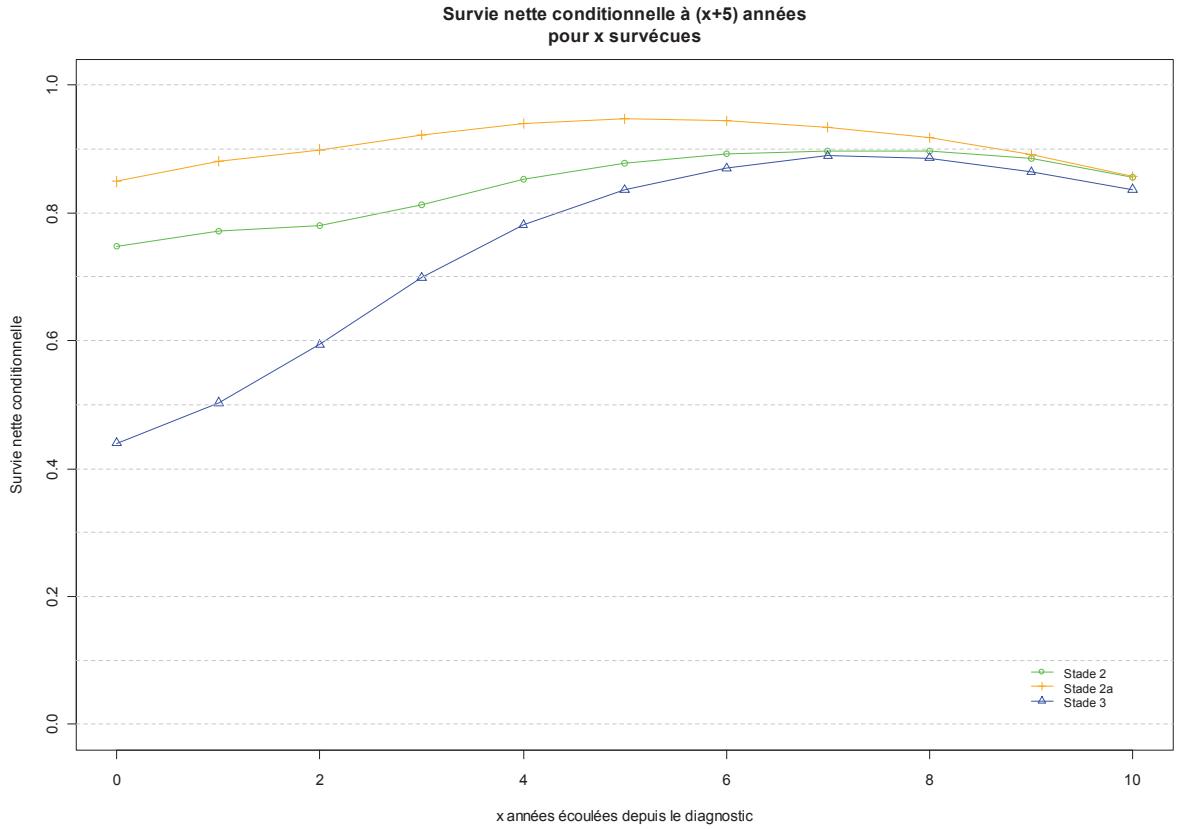
De manière générale, les résultats sont en faveur d'une augmentation de la survie à (x+5) ans conditionnellement au fait d'avoir déjà survécu un certain nombre x d'années, en tout cas pour les patients âgés de moins de 75 ans, ce qui est cohérent avec le profil de diminution avec le temps du taux de mortalité en excès.

La question posée initialement était, rappelons-le, d'examiner si l'effet du stade était constant ou non au cours du temps. Sous l'hypothèse d'un effet constant, la différence de taux en excès et donc de survie nette conditionnelle devrait se maintenir dans le temps entre les deux stades. La figure IV.2 ne va pas à l'évidence dans ce sens, pas plus que les résultats de la table IV.6 qui montrent des différences s'amenuisant avec le temps. Cependant, ces résultats sont assez indirects et pour illustrer directement l'effet du stade en fonction du temps, il nous faut examiner l'évolution du taux relatif associé au stade avec le temps, objet des figures IV.4 et IV.5.

Table IV.6. Survie nette conditionnelle à (x+5) ans sachant que le patient a déjà vécu x = 1, x = 3, x = 5 ou x = 10 ans

	Survie nette conditionnelle (années après le diagnostic)			
	1	**3**	**5**	**10**
Stade 2	**77.2 [70.0 ; 82.1]**	**81.4 [72.9 ; 86.7]**	**87.7 [77.4 ; 92.8]**	**85.5 [67.1 ; 93.5]**
[15 ;45]	78.4 [59.5 ; 87.7]	83.9 [66.0 ; 92.3]	88.4 [67.7 ; 95.6]	96.7 [77.5 ; 99.5]
[45 ;55]	78.2 [66.7 ; 85.0]	83.1 [72.2 ; 89.5]	88.2 [77.2 ; 93.7]	95.4 [83.6 ; 98.7]
[55 ;65]	77.7 [71.1 ; 82.4]	82.2 [75.1 ; 87.0]	88.0 [80.7 ; 92.4]	93.3 [83.3 ; 97.1]
[65 ;75]	77.3 [71.0 ; 82.3]	81.4 [73.9 ; 86.8]	87.8 [78.6 ; 93.0]	89.3 [76.0 ; 95.4]
[75 ;++]	76.6 [64.0 ; 84.6]	80.5 [65.7 ; 88.7]	87.6 [71.5 ; 94.2]	75.8 [40.3 ; 92.4]
Stade 2a	88.1 [72.0 ; 93.5]	92.3 [72.0 ; 96.9]	94.7 [74.6 ; 98.8]	85.7 [62.1 ; 95.4]
[15 ;45]	86.1 [54.6 ; 95.5]	91.5 [59.9 ; 98.1]	95.8 [66.3 ; 99.4]	98.2 [67.7 ; 99.9]
[45 ;55]	86.8 [68.6 ; 94.1]	91.8 [72.4 ; 97.2]	95.6 [75.6 ; 99.1]	96.9 [72.4 ; 99.6]
[55 ;65]	87.6 [75.3 ; 93.1]	92.1 [77.8 ; 96.7]	95.1 [80.6 ; 98.8]	93.5 [75.9 ; 98.4]
[65 ;75]	88.2 [72.5 ; 94.5]	92.4 [73.2 ; 97.4]	94.7 [75.5 ; 98.9]	88.1 [65.2 ; 96.4]
[75 ;++]	89.0 [61.2 ; 96.5]	92.5 [56.9 ; 98.6]	93.9 [57.6 ; 99.3]	72.5 [23.0 ; 94.6]
Stade 3	**50.3 [42.2 ; 56.4]**	**69.9 [56.9 ; 77.5]**	**83.6 [64.5 ; 90.0]**	**83.7 [14.2 ; 93.2]**
[15 ;45]	84.1 [40.0 ; 94.1]	85.2 [41.4 ; 94.7]	90.7 [52.6 ; 97.4]	100.0 [0.00 ; 100]
[45 ;55]	53.2 [38.3 ; 64.6]	60.9 [42.5 ; 72.5]	75.2 [50.3 ; 86.5]	99.5 [0.00 ; 100]
[55 ;65]	47.0 [37.0 ; 55.5]	61.9 [49.8 ; 71.3]	77.4 [60.9 ; 85.9]	98.6 [0.00 ; 99.9]
[65 ;75]	50.2 [40.1 ; 59.5]	70.3 [55.9 ; 79.8]	84.3 [65.0 ; 91.8]	96.6 [0.00 ; 99.5]
[75 ;++]	48.6 [34.1 ; 60.6]	76.3 [53.9 ; 87.7]	88.5 [51.3 ; 96.7]	59.1 [14.5 ; 89.4]

Figure IV.3. Survie nette conditionnelle au cours du temps par stade



IV.3.3. Taux relatifs

La figure IV.4 et la figure IV.5 présentent l'évolution du taux relatif respectivement entre le stade 2 et le stade 3 et entre le stade 2a et stade 3 au cours du temps tout d'abord tous âges confondus puis par classe d'âge. Nous pouvons observer qu'en prenant comme référence le stade 2 ou stade 2a, les différents taux relatifs ont tendance à diminuer au cours du temps.

Pour tous âges confondus, le taux relatif entre le stade 3 et le stade 2 diminue au cours du temps pour osciller autour de la valeur 1 à partir de 10 ans. Cependant, l'intervalle de confiance du taux relatif inclut la valeur 1 dès environ 5 ans après le diagnostic. Nous pouvons donc dire que nous ne pouvons plus montrer de différence significative entre les patients diagnostiqués en stade 2 et les patients diagnostiqués en stade 3 dès (approximativement) 5 ans de suivi après le diagnostic. Ces observations sont également valables pour le taux relatif entre stade 3 et stade 2a. En effet, les valeurs du taux relatif dans ce cas sont supérieures au taux relatif entre le stade 3 et le stade 2 mais l'intervalle de

confiance du taux relatifs inclut également la valeur 1 entre 4 et 5 ans après le diagnostic. De la même façon, nous pouvons dire que nous ne pouvons plus montrer de différence significative entre les patients diagnostiqués en stade 2a et les patients diagnostiqués en stade 3 approximativement entre 4 et 5 ans de suivi après le diagnostic.

En ce qui concerne la classe d'âge la plus jeune, nous pouvons observer que le taux relatif stade 3/stade 2 est proche de 1 jusqu'à 8 ans avec une petite tendance à un taux relatif inférieur à 1, c'est-à-dire défavorable pour les stades 2 (ce qui confirme bien les résultats table IV.5 et figure IV.2). Ce taux relatif diminue ensuite vers des valeurs extrêmes en fin de suivi. Ceci est dû au fait que le taux de mortalité des patients les plus jeunes diagnostiqués en stade 3 atteint des valeurs extrêmement faible en fin de suivi (figure IV.2). L'intervalle de confiance inclut la valeur 1 dès le début du suivi, la différence de survie entre les patients diagnostiqués en stade 2 et les patients diagnostiqués en stade 3 pour la classe d'âge [15 ;45] n'est donc pas significative tout au long du suivi. En ce qui concerne le taux relatif stade 3/stade 2a pour cette classe d'âge, il est inférieur à 1 jusqu'à 3 ans de suivi puis devient positif pour ensuite diminuer vers des valeurs extrêmes à partir de 9 ans. Tout comme le taux relatif stade 3/stade 2, l'intervalle de confiance inclut la valeur 1 dès le début du suivi, la différence de survie entre les patients diagnostiqués en stade 2a et les patients diagnostiqués en stade 3 pour la classe d'âge [15 ;45] n'est donc pas significative tout au long du suivi. La tendance du taux relatif à être inférieur est inattendu mais cette classe d'âge étant de taille très faible, il est difficile d'exploiter d'avantage les données.

Jusqu'à 9 ans de suivi les patients diagnostiqués en stade 2 de la classe d'âge [45 ;55] ont une meilleure survie que les patients diagnostiqués en stade 3 (respectivement 10 ans de suivi pour les patients diagnostiqués en stade 2a). Cette tendance s'inverse après. Le taux relatif atteint des valeurs extrêmes en fin de suivi pour la même raison qu'expliquée dans le paragraphe précédent. Nous pouvons voir que jusqu'à 7 ans, l'intervalle de confiance est proche de la valeur 1 mais ne l'englobe pas. La différence entre les stades n'est plus significative à partir de 7 ans.

Pour la classe d'âge [55 ;65], le taux relatif à la même tendance que pour la classe d'âge [45 ;55], la différence entre les stade étant légèrement plus marquée. L'intervalle de confiance inclut la valeur 1 entre 6 et 7 ans pour le taux relatif stade 3/stade 2 et entre 7 et 8 ans pour le taux relatif stade 3/stade 2a. Cela signifie qu'à partir de ce moment-là, il n'y a plus de différence significative entre les patients diagnostiqués en stade 2 (respectivement stade 2a) et les patients diagnostiqués en stade 3.

Pour la classe d'âge [65 ; 75], le taux relatif stade 3/stade 2 est supérieur à 1 jusqu'à environ 9 ans puis devient inférieur à 1 ensuite. Cependant, en observant l'intervalle de confiance, celui-ci inclut la valeur 1 entre 4 et 5 ans de suivi. Il n'y a donc plus de différence significative entre les patients diagnostiqués en stade 2 et les patients diagnostiqués en stade 3 lorsque 4 à 5 ans de suivi se sont écoulés. En ce qui concerne le taux relatif stade 3/stade 2a, la valeur 1 est atteinte entre 10 et 11 ans. Cependant, en observant l'intervalle de confiance, celui-ci inclut la valeur 1 entre 4 et 5 ans de suivi. Il n'y a donc également plus de différence significative entre les patients diagnostiqués en stade 2a et les patients diagnostiqués en stade 3 à partir de la période de 4 à 5 ans de suivi.

Enfin, en ce qui concerne la classe d'âge la plus âgée, la différence entre les stades 2 et les stades 3 est marquée en début de suivi puis s'atténue. L'intervalle de confiance inclut la valeur 1 entre 3 et 4 ans de suivi : il n'y a donc plus de différence significative à partir de ce moment-là entre les stades. En ce qui concerne le taux relatif stade 3/stade 2a, il reste supérieur à 1 tout au long du suivi, les stades 3 sont donc plus à risque que les stades 2a. Cependant, l'intervalle de confiance inclut la valeur 1 entre 3 et 4 ans de suivi. Il n'y a donc plus de différence significative entre les patients diagnostiqués en stade 2a et les patients diagnostiqués en stade 3 à partir de cette période.

Figure IV.4. Taux Relatifs Stade3/Stade2

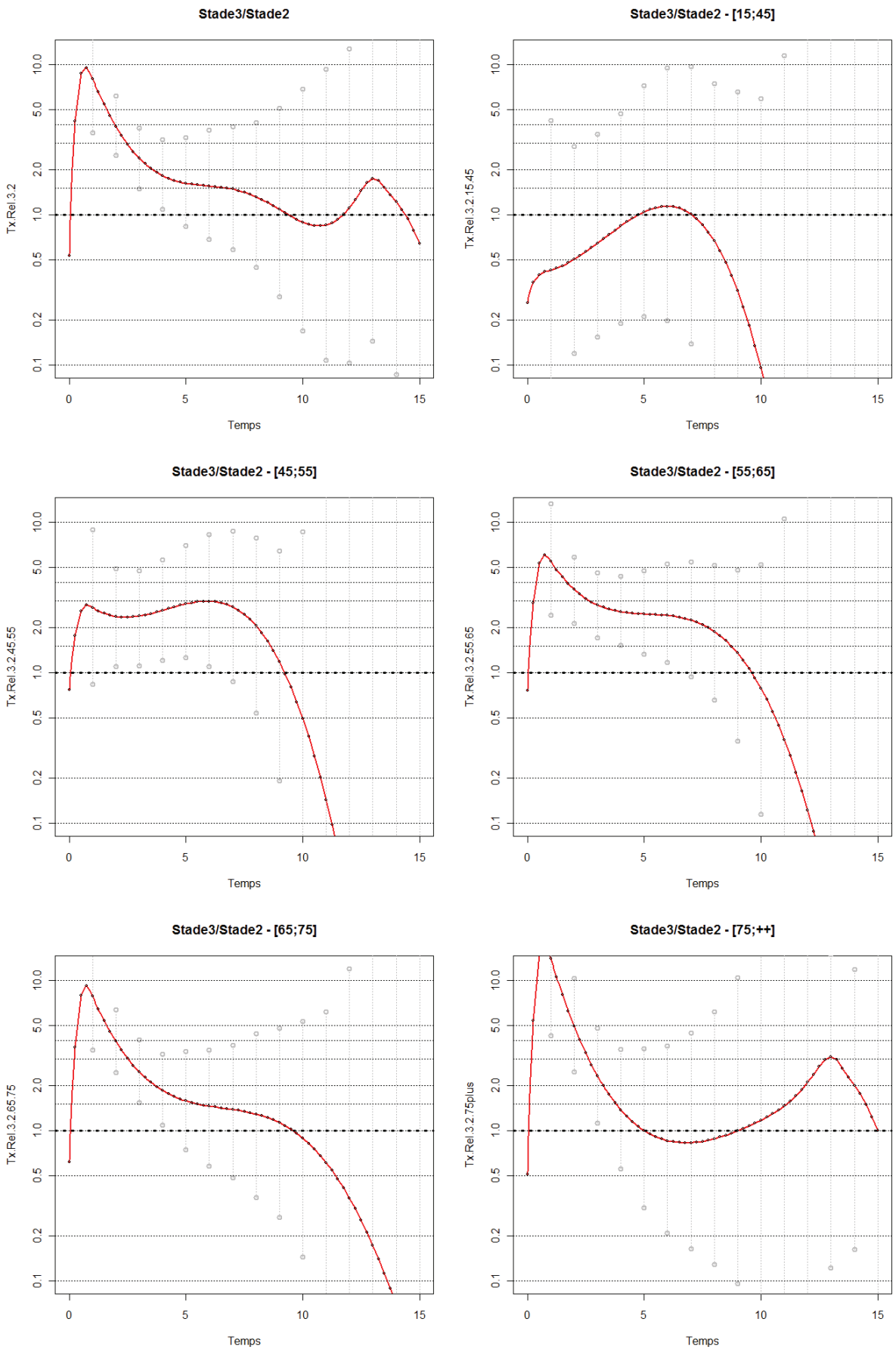
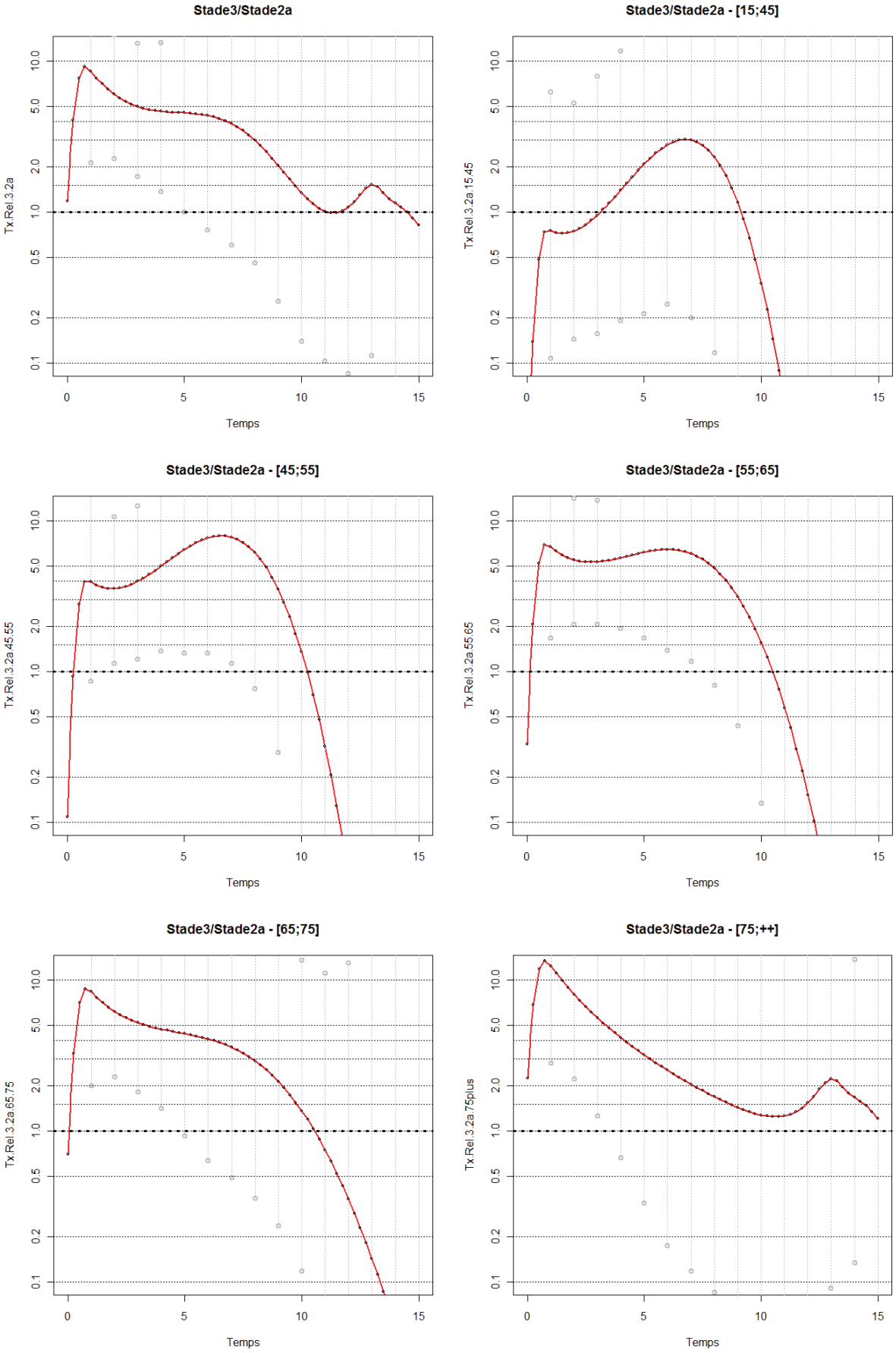


Figure IV.5. Taux Relatifs Stade3/Stade2a



IV.4. Discussions

Cette étude a permis l'estimation à long terme, encore peu répandue, de la survie nette conditionnelle par stade pour le cancer du Côlon.

Globalement, d'après les résultats, l'effet du stade sur le taux de mortalité en excès diminue au cours du temps. Les taux relatifs stade 3/stade 2 et stade 3/stade 2a (figure IV.4 - figure IV.5) montre que ces derniers se différencient jusqu'à environ 10 ans de suivi après le diagnostic, la différence étant non significative à partir de 5 ans de suivi environ, puis atteignent un même ordre de grandeur. Le graphe de la survie nette conditionnelle (figure IV.3) confirme cette observation : en fin de suivi, les courbes obtenues pour les différents stades analysés sont très proches. Cependant, certains résultats nous interpellent. D'une part, nous pouvons remarquer que ni pour le stade 2, ni pour le stade 3, un palier est atteint en ce qui concerne la survie nette. Ceci est surprenant, notamment pour les stades 2 pour lesquels nous nous attendions à ce qu'il n'y ait plus de décès en excès à long terme. Différentes études ont montré qu'il pouvait y avoir des problèmes de classification si le nombre de ganglions examinés est inférieur à 8 ; certains ganglions peuvent être envahis mais non examinés, le stade est alors sous-évalué (les stades 2 sont alors un mélange de stade 2 et de stades plus avancés). C'est pourquoi nous nous sommes intéressés aux patients diagnostiqués en stade 2 avec plus de 8 ganglions, appelé stade 2a, qui sont en réalité les « vrais » stades 2. Cependant, nous pouvons voir également que la courbe des stades 2a n'atteint pas de palier non plus, ce qui est un résultat surprenant. Afin de voir si ces tendances n'étaient pas dues à la modélisation utilisée, les données ont été analysées également avec des natural splines afin de contraindre les taux de mortalité à être linéaire en dehors des nœuds et d'éviter de potentiels effets de bord. Les résultats ont montré les mêmes tendances, ils ne semblent donc pas sensibles à la modélisation.

D'après [VanSteenBergen, 2013], le seuil de mortalité en excès négligeable est atteint lorsque la survie nette conditionnelle atteint 95%. Ce seuil signifie qu'une fois que la survie nette conditionnelle a atteint les 95%, la survie du groupe de patients est quasiment comparable à celle d'un groupe de patients de la population générale ayant les mêmes caractéristiques démographiques. Pour les patients diagnostiqués en stade 2, ce seuil est atteint pour les classes d'âges les plus jeunes ([15 ; 45[et [45 ; 55[) entre 10 et 15 ans de suivi après avoir déjà survécu respectivement 5 et 10 ans depuis le diagnostic. Pour les patients diagnostiqués en stade 2a, ce seuil est atteint entre 8 et 10 ans de suivi après avoir déjà

survécu respectivement 3 et 5 ans depuis le diagnostic pour les classes d'âges suivantes : [15 ; 45[, [45 ; 55[, [55 ; 65[. Pour les patients diagnostiqués en stade 3, ce seuil est atteint entre 10 et 15 ans de suivi après avoir déjà survécu 5 et 10 ans depuis le diagnostic pour toutes les classes d'âges sauf la plus âgées [75 ; ++[. Ces derniers résultats sont à prendre avec précautions comme peut nous l'indiquer l'étendue des intervalles de confiance.

Lorsque nous regardons les quantiles de décès, 90% des décès pour les stades 2 ont eu lieu pendant les 14 premières années de suivi alors qu'en ce qui concerne le stade 3, 90% des décès ont eu lieu pendant les 10 premières années de suivi. Cela signifie qu'il ne reste plus beaucoup d'information à long terme, particulièrement pour les stades avancés. Il est donc très difficile de comparer les taux en excès à long terme entre deux populations telles que les stades 2 et les stades 3. Les estimations des taux relatifs présentent également une grande variabilité à long terme. Ceci amène à penser que la méthodologie utilisée n'est pas adéquate et qu'une réflexion méthodologique supplémentaire est nécessaire pour résoudre ce problème. Comme évoqué au début de la discussion, l'utilisation des natural splines n'a pas changé les résultats, nous nous demandons alors su nous devons garder notre outil pour modéliser le taux en excès. Si oui, peut-être devons-nous nous interroger sur la variabilité du critère d'Akaike pour le choix du modèle [Commenges, 2008] ainsi que sur l'inclusion de la covariable stade dans le modèle. Une autre proposition serait de modéliser les données à partir de 5 ans de suivi afin d'avoir une dynamique plus simple à modéliser. Des recherches à ce sujet sont donc nécessaires dans le futur.

L'estimation de la survie nette à long terme est délicate d'autant plus lorsque l'on se situe dans la classe la plus âgée. En effet, les personnes âgées ont un risque plus élevé de décéder du cancer ainsi que de décéder des autres causes, ce qui peut perturber le traitement, les soins et le rétablissement. La question est la suivante : l'estimation à long terme de la survie nette chez les personnes âgées a-t-elle un sens ? Cette estimation de la survie nette à long terme pour les patients âgés est donc très discutable.

Pour conclure, la différence de survie entre les patients diagnostiqués en stade 2 et ceux diagnostiqués en stade 3 n'est plus significative après avoir survécu 4 ans depuis le diagnostic. L'effet du stade s'atténue donc au cours du temps. Par ailleurs, les résultats tous âges confondus montrent un excès de mortalité persistant à distance du diagnostic. Cependant, les résultats par âge suggèrent une atténuation notable de cet excès de mortalité chez les patients les plus jeunes. Cette observation a pu être traduite en terme de survie conditionnelle, cette dernière atteignant en effet à 15 ans (conditionnellement au fait d'avoir survécu 10 ans)

des valeurs proches de 95%. L'imprécision de ces estimations constitue une limite majeure de notre étude, notamment pour les stades 3 pour lesquels elle rend toute interprétation impossible. La survie conditionnelle, sous réserve de données a long terme suffisantes, représente un indicateur intéressant. Il permet , tout comme l'évolution du taux de mortalité en excès, d'envisager les stratégies de suivi les plus appropriées. Les épidémiologistes et les cliniciens semblent trouver la survie conditionnelle plus "parlante" que l'évolution du taux en excès , peut-être parce que la notion de survie leur est plus habituelle.

Chapitre V

Conclusion et perspectives

V.1. Conclusion

Dans le cadre de ce travail, nous nous sommes intéressés à l'estimation de la survie nette suite au diagnostic d'un cancer, c'est-à-dire, la survie que l'on observerait dans le cadre hypothétique où la seule cause de mortalité possible était le cancer. L'intérêt principal de cet indicateur épidémiologique est qu'il permet la comparaison entre pays ou entre périodes de diagnostic en termes de survie et permet donc une comparaison directe de l'efficacité et de l'amélioration des systèmes de soins concernant le cancer.

Parmi les différentes méthodes développées dans le but d'estimer la survie nette, deux méthodes ont été retenues suite à une étude de simulation qui a fait l'objet du chapitre II de cette thèse. Le modèle paramétrique multivarié et la méthode non-paramétrique de Pohar-Perme sont les seules méthodes capables de fournir des estimations non biaisées de la survie nette. Ces deux méthodes prennent en compte la censure informative de différentes manières : la méthode paramétrique ajuste le modèle sur les covariables induisant une censure informative et la méthode non-paramétrique utilise la pondération par l'inverse de la probabilité de censure. L'utilisation de l'une ou de l'autre de ces deux méthodes dépend de l'objectif de l'analyse. Si l'objectif est d'obtenir une estimation ponctuelle de la survie avec son intervalle de confiance, la méthode non-paramétrique de Pohar-Perme est recommandée. En revanche, si l'objectif est de quantifier les effets des variables pronostiques agissant sur le taux de mortalité en excès, le modèle paramétrique multivarié est recommandé.

Ces résultats ont été publiés dans une revue statistique, mais il était important de les faire connaître à la communauté épidémiologique. Le choix a donc été fait de publier un article, dans une revue épidémiologique, destiné à illustrer l'impact de l'utilisation des méthodes anciennes plutôt que la méthode de Pohar-Perme quand l'objectif est de fournir des estimations ponctuelles de la survie nette à 5, 10 et 15 ans (non standardisées puis standardisées) à partir de données réelles. En considérant la méthode de Pohar-Perme comme gold-standard, les résultats ont montré que l'erreur était plus importante particulièrement pour

les cancers de bon pronostic (mortalité globale d'avantage influencée par la mortalité autres-causes impliquant une augmentation de la censure informative liée à ces décès), pour des temps de suivi plus long (population de plus en plus déformée). Ces résultats ont fait polémique et cet article a reçu quelques critiques dans le fait de promouvoir la méthode de Pohar-Perme au détriment des autres méthodes existantes.

L'inconvénient majeur du modèle paramétrique du taux en excès est qu'il nécessite une stratégie de construction de modèle qui s'avère complexe et difficile à mettre en place. Cette stratégie nécessite, dans un premier temps, de faire des hypothèses pour l'élaboration du modèle. Dans un second temps, elle nécessite d'évaluer si les données sont bien reflétées par le modèle. Pour vérifier chacune des hypothèses faites antérieurement dans le modèle, nous avons proposé différents tests issus d'un même cadre théorique, les transformées de martingale, permettant de vérifier l'hypothèse des taux proportionnels, ainsi que la forme fonctionnelle et la fonction de lien utilisées dans le modèle. Ces différents outils constituent un test formel pour chacune des hypothèses ainsi qu'une représentation graphique permettant de voir où se situe approximativement le caractère non-proportionnel en fonction du temps, le caractère non-linéaire en fonction de la covariable d'intérêt ou la mauvaise spécification de la fonction de lien en fonction de logarithme du taux relatif. Les résultats de ce travail seront soumis à la revue Statistics in Medicine en fin d'année.

Enfin, une application épidémiologique visant à étudier l'impact des facteurs pronostiques, tel que le stade au diagnostic, sur la dynamique du taux de mortalité en excès après la survenue d'un cancer du côlon a été effectuée. Cette étude a permis de montrer qu'à partir de 5 ans environ, nous ne pouvons plus montrer de différence significative entre les patients diagnostiqués en stade 2 et les patients diagnostiqués en stade 3 après le diagnostic en terme de taux de mortalité en excès. Elle a permis d'objectiver également que le taux de mortalité en excès pour certains stades et certains âges atteignaient une valeur très faible, pouvant être considérée comme négligeable au-delà d'un certain délai. La survie nette conditionnelle, qui rejoint l'analyse de la dynamique du taux de mortalité en excès, a pu être fournie. Cet indicateur, qui consiste à restituer la survie à 5 ans (par exemple) conditionnellement au fait d'avoir survécu 1, 2, 3,.., 5 ans, est de plus en plus restitué dans les revues épidémiologiques, alors qu'il contient la même information que l'évolution du taux de mortalité en excès, à savoir qu'une survie conditionnelle qui augmente avec le temps n'est que le reflet d'un taux de mortalité en excès allant en diminuant. Il semble cependant que son

utilisation soit perçue comme plus facilement et directement utilisable par les cliniciens pour décider des stratégies de suivi des patients.

V.2. Perspectives

Nous présentons dans ce paragraphe des perspectives de recherche et de développements selon les différents axes de travail explorés dans le cadre de cette thèse.

Tests des hypothèses d'un modèle paramétrique du taux en excès

Le but principal du chapitre III de cette thèse est de proposer une boîte à outils permettant de tester les principales hypothèses d'un modèle de taux en excès dans un même cadre théorique. Les résultats présentés dans ce chapitre ne portent que sur le test de l'hypothèse des taux proportionnels. L'implémentation des tests portant sur la forme fonctionnelle et sur la fonction de lien est en cours. L'analyse de performance de ces deux tests sera faite une fois la phase d'implémentation terminée selon les mêmes modalités d'évaluation que le test de l'hypothèse des taux proportionnels.

Une question que nous nous posons est l'intérêt d'un test « omnibus » pour tester l'adéquation globale du modèle. En effet, ce test permettrait de tester l'adéquation globale du modèle, c'est-à-dire, de rejeter ou non l'hypothèse nulle H_0 avec H_0 : « le modèle est correcte », sans se focaliser sur une particularité de ce modèle telle que le caractère proportionnel de l'effet des covariables, la forme fonctionnelle des covariables ou la fonction de lien. Si l'hypothèse nulle n'est pas rejetée, le modèle est considéré comme correct, sinon, il est considéré comme incorrect et dans ce cas, des analyses supplémentaires portant sur les différentes spécificités du modèle sont alors nécessaires pour évaluer celle(s) qui ne correspondent pas aux données.

Cependant, la performance des tests d'une spécificité particulière peut être impactée par une mauvaise spécification des autres aspects de la modélisation de la covariable d'intérêt (comme nous avons pu le voir par exemple, la non prise en compte du caractère non-linéaire de l'effet d'une covariable peut impacter les performances du test testant l'hypothèse du caractère proportionnel de son effet). Les résultats obtenus sur données réelles doivent donc être interprétés avec précaution. Notamment, comme il a été recommandé [Abrahamovicz,

2007], le test de l'hypothèse des taux proportionnel doit être effectué en spécifiant de manière systématique un effet non-linéaire pour la covariable d'intérêt.

Pour remédier à ce problème, une solution serait de développer un outil permettant de tester de façon *simultanée* les différentes hypothèses du modèle, notamment les plus importantes, à savoir la proportionnalité et la forme fonctionnelle. Il permettrait de se rendre compte si seulement l'une d'entre elle est mal spécifiée ou les deux (cet outil ne reposerait pas sur l'hypothèse « toutes les autres hypothèses sont correctes »). Différents outils [Sasieni, 2003][Pohar, 2008] ont été développés dans le cadre de la survie globale. Une perspective intéressante mais complexe pourrait être d'adapter ces outils au cadre de la survie nette.

Lorsque ces outils auront été implémentés et testés (tous ou partiellement), une autre perspective pourrait être de les rendre disponibles en routine, via un package R afin que l'étape de vérification des hypothèses du modèle, qui sera alors rapide, soit accessible à tous.

Outils de modélisation de la survie à long terme

L'étude de l'impact à long terme des facteurs pronostiques a présenté des difficulté importantes qui n'avaient pas été anticipées au début de cette thèse. En effet, nous avons pu remarquer tout d'abord le problème du manque d'information à long terme, notamment pour les stades les plus avancés et pour la classe d'âge la plus âgée (qui est la plus représentée pour le cancer du côlon). Ceci conduit à des estimations de survie présentant des intervalles de confiance très grands. Ce manque d'information a aussi été rencontré pour le stade le moins avancé pour lequel les modèles avaient du mal à converger du fait d'un très faible taux de mortalité en excès.

Afin de pallier au manque d'information liée à une analyse stratifiée par stade, il pourrait être envisagé de garder la méthodologie utilisée et d'inclure la covariable stade comme une variable catégorielle dans le modèle. La principale difficulté de cette « solution » réside dans la complexité du modèle à construire, sachant qu'en plus de la covariable âge, il faudra modéliser un effet pour chaque stade et tester toutes les combinaisons possibles des fonctions candidates pour chaque effet. Afin de limiter cette complexité, une solution serait de modéliser les données à partir de 5 ans de suivi afin d'avoir une dynamique plus simple à modéliser.

En conclusion, cette thèse a contribué à mieux connaître les propriétés des différents estimateurs de la survie nette. De plus, un cadre théorique a été proposé afin, qu'à terme, il soit mis à la disposition des épidémiologistes des outils permettant de vérifier les hypothèses des modèles qu'ils utilisent.

Bibliographie

Abrahamowicz M, MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in Medicine*. 2007 Jan 30; 26(2):392-408.

Akaike H. Information theory and an extension of the maximum likelihood principle. *In 2nd International Symposium on Information Theory*, Petrov BN, Csaki F (eds). 1973; 268–281.

Altekruse SF, Kosary CL, Krapcho M, Neyman N, Aminou R, Waldron W, Ruhl J, Howlader N, Tatalovich Z, Cho H, Mariotto A, Eisner MP, Lewis DR, Cronin K, Chen HS, Feuer EJ, Stinchcomb DG, Edwards BK (eds). *SEER Cancer Statistics Review*, 1975-2007, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2007/, based on November 2009 SEER data submission, posted to the SEER web site, 2010. Accessed October 16, 2014.

Ashworth TG. Inadequacy of death certification: proposal for change. *Journal of Clinical Pathology*. 1991; 44(4):265–268.

Barlow WE, Prentice RL. Residual for relative risk regression. *Biometrika*. 1988, Mar;75(1):65-74

Berkson J, Gage RP. Calculation of survival rates for cancer. *Proceeding of the Staff Meeting of the Mayo Clinic*. 1950; **25**: 270-86.

Binder-Foucard F, Belot A, Delafosse P, Remontet L, Woronoff AS, Bossard N. Estimation nationale de l'incidence et de la mortalité par cancer en France entre 1980 et 2012. Partie 1 - Tumeurs solides. Saint-Maurice (Fra) : Institut de veille sanitaire, 2013. 122 p.

Böhmer PE. Theorie der unabhängigen Wahrscheinlichkeiten Rapports. Mémoires et procès verbaux du septième congrès international d'actuaires. Amsterdam. 1912; 2:327-43.

Bossard N, Velten M, Remontet L, et al. Survival of cancer patients in France: a population-based study from The Association of the French Cancer Registries (FRANCIM). *European Journal of* Cancer. 2007;43:149-60.

Commenges D, Sayyareh A, Letenneur L, Guedj J, Bar-Hen A. Estimating a difference of Kullback–Leibler risks using a normalized difference of AIC. *The annals of applied statistics*. 2008. 2(3) 1123-1142.

Cortese G, Scheike TH. Dynamic regression hazards models for relative survival. *Statistics in Medicine*. 2008 Aug 15;27(18):3563-84.

Cox, DR. Regression models and life tables. *Journal of the Royal Statistical Society, Series B*. 1972; 34:187–220.

Cox DR. Partial likelihood. *Biometrika*. 1975; 62:269_276.

Danieli C, Remontet L, Bossard N, Roche L, Belot A. Estimating net survival: the importance of allowing for informative censoring. *Statistics in Medicine*. 2012; 31:775–86.

De Angelis R, Francisci S, Baili P, et al. The EUROCARE-4 database on cancer survival in Europe: data standardisation, quality control and methods of statistical analysis. *European Journal of* Cancer. 2009; 45(6):909–30.

De Angelis R, Sant M, Coleman MP, Francisci S, et al. Cancer survival in Europe 1999–2007 by country and age: results of EUROCARE-5 - a population-based study. *The lancet oncology 2014; 15*(1):23-34.

Dickman PW, Lambert PC, Coviello E, Rutherford MJ. Estimating net survival in population-based cancer studies. *International Journal of Cancer*. 2013; (133)2 :519-521,

Ederer F, Heise H. Instructions to ibm 650 programmers in processing survival computations, methodological note no. 10, end results evaluation section. *Technical report, National Cancer Institute, Bethesda MD*. 1959.

Ederer F, Axtell LM, Cutler SJ. The relative survival rate: a statistical methodology. *National Cancer Institute Monograph* 1961; 6:101-21.

Esteve J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine* 1990; 9(5):529–538.

Fleming TR, Harrington DP. Counting processes and survival analysis. New York: Wiley. 1991.

Grambsch PM, Therneau TM. Proportional hazard tests and diagnostics based on weighted residuals. *Biometrika*. 1994 Aug; 81 (3): 515-526.

Hakulinen T, Tenkanen L. Regression analysis of relative survival rates. *Applied Statistics* 1987; 36:309–317.

Howlader N, Noone AM, Krapcho M, Garshell J, Miller D, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z,Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). *SEER Cancer Statistics Review*, 1975-2011, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2011/, based on November 2013 SEER data submission, posted to the SEER web site, April 2014. Accessed October 16, 2014.

Kaplan EL, Meier P. Non parametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958 Jun; 53: 457-481.

Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*. 1993 Sep. 80; (3): 557-572.

Lin DY, Spiekerman CF. Model checking techniques for parametric regression with censored data. *Scandinavian Journal of Statistics*. 1996; 23: 157-177.

Meyer P. A Decomposition theorem for supermartingales. *Illinois Journal of Mathematics*. 1962; 6: 193–205.

Meyer P. Decomposition of supermartingales: the uniqueness theorem. *Illinois Journal of Mathematics*. 1963; **7**: 1–17

Percy C, Stanek E, Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. *American Journal of Public Health*. 1981 Mar; 71(3):242–250.

Perme MP. Goodness of fit of relative survival models. 2007

Perme MP, Andersen PK. Checking hazard regression models using pseudo-observations. *Statistics in Medicine*. 2008 Nov 10; 27(25):5309-28.

Perme MP, Stare J, Estève J. On estimation in relative survival. *Biometrics*. 2012 Mar; 68(1):113-20.

Remontet L, Bossard N, Belot A, Esteve J. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in Medicine*. 2007 May 10; 26(10):2214-2228.

Robins, J. M. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the American Statistical Association-Biopharmaceutical Section*. 1993. pp. 24-33. Alexandria, Virgnia, U.S.

Roche L, Danieli C, Belot A, Grosclaude P, Bouvier AM, Velten M, Iwaz J, Remontet L, Bossard N. Cancer net survival on registry data: use of the new unbiased Pohar-Perme estimator and magnitude of the bias with the classical methods. *International Journal of Cancer*. 2013 May 15; 132(10):2359-69.

Roche L, Danieli C, Belot A, Iwaz J, Remontet L, Bossard N. Author's reply to: Estimating net survival in population-based cancer studies. *International Journal of Cancer*. 2013 July; (133)2: 522-523

Ross SM. Simulation, Fourth Edition. Elsevier Academic Press: Amsterdam, 2006.

Sasieni PD. Proportional excess hazards. *Biometrika* 1996; 83(1):127–141.

Sasieni PD. Martingale difference residuals as a diagnostic tool for the Cox model. *Biometrika*. 2003. 90:899-912.

Satten, G. A., Datta, S., and Robins, J. Estimating the marginal survival function in the presence of time dependent covariates. *Statistics and Probability Letters*. 2001. **54,** 397–403.

Schoenfeld D. Partial residuals for the proportional hazards regression model. *Biometrika*. 1982 Apr. 69 (1): 239-241.

Stare J, Pohar M, Henderson R. Goodness of fit of relative survival models. *Statistics in Medicine*. 2005 Dec 30; 24(24):3911–3925.

Sylvestre M.-P., Abrahamowicz M. Comparisons of algorithm to generate event times conditional on time-dependent covariates. *Statistics in Medicine*. 2008; 27:2618-2634.

Therneau T. M., Grambsch P. M., Fleming T. R. Martingale-Based residuals for survival models. *Biometrika*. 1990 Mar. 77 (1): 147-160.

Van Steenbergen LN, Steur M, Lemmens VE, Rutten HJ, Van Spronsen DJ, Janssen-Heijnen ML. Minimal excess mortality for long-term colon cancer survivors in the Netherlands 1989-2008. *European Journal of Cancer*. 2013 Feb;49(3):585-92.

Résumé
La survie nette est un indicateur très utilisé en épidémiologie des cancers. Il s'agit de la survie que l'on observerait si la seule cause de mortalité était le cancer ; il est le seul indicateur épidémiologique utilisable à des fins de comparaisons de survie (entre périodes/pays) car il s'affranchit des éventuelles différences de mortalité dues aux autres causes que le cancer.
Le premier objectif de notre travail était d'analyser les performances des différentes méthodes d'estimation de la survie nette sur données simulées ainsi que sur données réelles afin que les méthodes non biaisées soient reconnues scientifiquement et soient les seules à être utilisées par la suite. Nous avons ainsi démontré que deux approches étaient capables d'estimer sans biais la survie nette : l'approche non paramétrique de Pohar-Perme et l'approche reposant sur une modélisation multivariée du taux de mortalité en excès dû au cancer. Cette dernière approche impose une stratégie de construction difficile à mettre en place.
Le deuxième objectif était de développer une boîte à outils composée de différents tests permettant de vérifier les différentes hypothèses faites lors de la construction d'un modèle de régression du taux de mortalité en excès. Ces hypothèses concernent habituellement la proportionnalité ou non de l'effet des covariables, leur forme fonctionnelle, ainsi que la fonction de lien utilisée.
Le troisième objectif était une application épidémiologique qui visait à étudier l'impact des facteurs pronostiques, tel que le stade au diagnostic, sur la survie nette conditionnelle, en d'autres termes sur la dynamique du taux de mortalité en excès, après la survenue d'un cancer du côlon.

Mots-clé : Survie nette; Censure informative; Modèle multivarié; Tests ; Proportionnalité ; Forme fonctionnelle, Fonction de lien; Survie nette conditionnelle; Stade au diagnostic ; Cancer

Title : Methodological contribution to net survival estimation: estimator comparison and test of the parametric hazard model assumption

Summary
Net survival is one of the most important indicators in cancer epidemiology. It is defined as the survival that would be observed if cancer were the only cause of death. This is the only one indicator allowing comparisons of cancer impact between countries or time periods because it is not influenced by death because of other causes.
The first objective of this work was to compare the performance of several estimators of the net survival in a simulation study and then on real data in order to promote unbiased methods. Those methods are the non-parametric Pohar-Perme method and the parametric multivariable excess rate model. The latest one needs a model building strategy.
The use of diagnostic procedures for model checking is an essential part of the modeling process. The second objective was to develop a tool box composed of diagnostic tools allowing to check hypothesis usually considered when constructing an excess mortality rate model, that is, the proportionality or not of the effect of covariates, their functional form and the link function.
The third objective deals with the study of the impact of prognostic variables, such as stage at diagnosis, on conditional net survival, that is, on the dynamic of the excess hazard mortality after the diagnosis of colon cancer.

Keywords : Net survival; Informative censoring; Multivariable model; Tests; Proportionality; Functional form; Link function; Conditional net survival; Stage at diagnosis; Cancer

Intitulé et adresse du laboratoire :
Laboratoire de Biométrie et Biologie Evolutive - Equipe Biostatistiques-Santé
Service de Biostatistique - Centre Hospitalier Lyon-Sud
165 chemin du Grand Revoyet - Bâtiment 4D - 69495 Pierre-Bénite